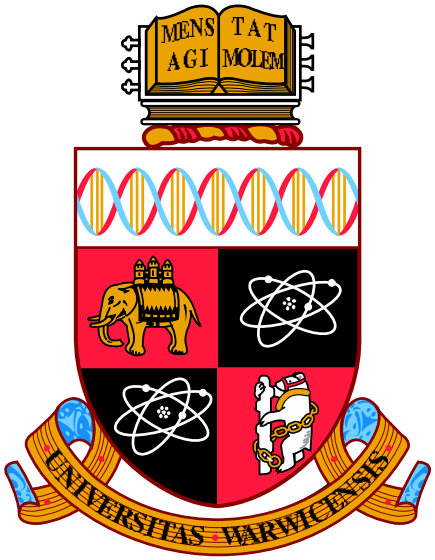


Stability and Biophysical Characterisation of Protein Therapeutics

Vol. 1

Meropi Sklepari

A thesis submitted to the University of Warwick for the degree of Doctor of Philosophy

Department of Chemistry

June 2017

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my project supervisors most sincerely: Prof. Alison Rodger and Dr. Claudia Blindauer of the Department of Chemistry, University of Warwick, for their constant advice and encouragement and for reminding me that I should persist in doing things right. Many thanks to key people for this work, especially Dr. Andrew Reason for his invaluable help, Michael-John O'Connor and Daniel Binet, of BioPharmaSpec Ltd, for passing on their expertise to me in the use of Mass Spectrometry and their invaluable guidance. Special thanks to Prof. Syma Khalid of University of Southampton for her support, Dr. Ivan Prokes of the Department of Chemistry, University of Warwick, for providing training in the use of NMR and for his technical advice.

I am more than grateful to my colleagues: Dr. Nikola Chmel for his major help in the lab and advice, Marco Pinto for his great contribution to my knowledge on FT-IR spectroscopy, the extremely long discussions and for our friendship, and Dr. Kasra Razmkhah for his help and the introduction to the lab. Many thanks to Dr. Shirin Jamshidi, Jacopo Franco and Dr. Claire Broughton for their contribution in some parts of this work. Thank you to the Marie Curie fellows especially Maria and Daniela for the discussions and friendship along the way.

I am extremely grateful to those who played a major role in my decision to start a PhD: Argy Vontzi, Prof. Panagiotis Marakos and Prof. Nikolais Pouli. I would like to express my warmest thanks to my dearest friends from Greece: Antigoni and Dimitra for their priceless support, Vasso for the endless cultural integration, Argyro and Margarita for bringing out my happy side, Apollon for his constant interest, Jorje and Matina for being a priceless connection between the UK and home. Finally, I would like to express my very deep gratitude to my beloved parents, mum and dad, for always being next to me full of belief, patience and love no matter what, and to my one and only brother, whose steps I followed.

Declaration

This thesis is submitted to the University of Warwick for the degree of Doctor of Philosophy. It has been composed by the author and has not been submitted in any previous application for any degree. The work presented (including data generated and data analysis) was carried out by the author, except where indicated in the text.

Abbreviations

ACS	Ammonium D-10-camphorsulfonate
ADC	Antibody drug conjugate
ATR FT-IR	Attenuated total reflectance Fourier transform infrared
BPCI	Biologics price competition and innovation
CBER	Centre for biologic evaluation and research
CD	Circular dichroism
CDER	Centre of drug evaluation and research
CKD	Chronic kidney disease
CDR	Complementary-determining regions
CHMP	Committee for medicinal products for human use
CRF	Chronic renal failure
DAR	Drug to antibody ratio
DC	Direct-current voltage
df	Dilution factor
DIT	Data integration time
DLS	Dynamic light scattering
DMARD	Disease-modifying antirheumatic drug
DNA	Deoxyribonucleic acid
DOSY	Diffusion-ordered spectroscopy
DSC	Differential scanning calorimetry
DSSP	Dictionary of protein secondary structure
EDTA	Ethylenediamine tetraacetic acid
EMA	European Medicines Agency
EPAR	European public assessment reports
ER	Endoplasmic reticulum
ES	Electrospray
ESA	Erythropoiesis-stimulating agent
ESI	Electrospray ionisation
FDA	Food and Drug Administration
FDC	Federal food, drug and cosmetic
FID	Free-induction decay signal
FWHH	Full width half height
FWHM	Full width half maximum
GABI	Generics and biosimilars initiative
GHD	Growth hormone deficiency
GlcNAc	N-acetylglucosamine
GMP	Good manufacturing practice
HC	Heavy chain

HER2	Human epidermal growth factor 2
hGH	Human growth hormone
HIV	Human immunodeficiency virus
HOMO	Highest occupied molecular orbital
HPLC	High-performance liquid chromatography
HPV	Human papilloma virus
HSA	Human serum albumin
HSQC	Heteronuclear single-quantum coherence
HT	High tension
ICH	International council on harmonisation
ICP-OES	Inductively coupled plasma optical emission spectrometer
Ig	Immunoglobulin
IGF	Insulin-like growth factor
IRE	Internal reflection element
LC	Light chain
LR	Lock-release
LUMO	Lowest unoccupied molecular orbital
mAb	Monoclonal antibody
MAH	Marketing authorisation holder
MALDI	Matrix assisted laser desorption ionisation
MD	Molecular Dynamics
MMAE	Monomethyl auristatin E
MRM	Multiple reaction monitoring
MRW	Mean residue weight
MS	Mass spectrometry
MW	Molecular weight
MWCO	Molecular weight cut-off
$Na[Co^{III}(EDDS)]$	Ethylenediaminedisuccinate
NMR	Nuclear magnetic resonance
NOESY	Nuclear Overhauser effect spectroscopy
NRMSD	Normalised residual mean square difference
PABC	p-aminobenzoylcarbamate
PBS	Phosphate-buffered saline
PDA	Photodiode array
PDB ID	Protein data bank identity
PDI	Polydispersity index
PHS	Public health service
PRCA	Pure red cell aplasia
PTMs	Post-translational modifications

Q-TOF	Quadrupole time-of-flight
QC	Quality control
RA	Rheumatoid arthritis
REMS	Risk evaluation and mitigation strategy
RF	Radiofrequency
RI	Refractive index
RT	Retention time
SA	Simulated annealing
SELCON	Self-consistent
SOP	Standard operating procedure
SSNN	Secondary structure neural network
t-PA	Tissue plasminogen activator
TIC	Total ion current
TNFR2	Tumour necrosis factor receptor 2
TOCSY	Total correlation spectroscopy
TQ-S	Triple quadrupole spectrometer
UPLC	Ultra-performance liquid chromatography
USPTO	United States patent and trademark office
UV-Vis	Ultraviolet-Visible
vcMMAE	Valine-citrulline monomethyl auristatin E
WHO	World health organisation

General Abstract

For the past two decades, the development of protein therapeutics has significantly expanded with numerous biopharmaceutical and biosimilar products entering the medicine market every year, and even more queuing in the pipeline globally. Biologics are very complex molecules and therefore extremely sensitive to minor changes in the manufacturing process, which can result in heterogeneity and affect the stability, potency and immunogenicity of the final product. Public health organisations, such as EMA (European Medicines Agency), require that biological products should be extensively tested for their similarity to the original drug (in the case of a biosimilar) as well as to products from different batches (batch-to-batch comparison). The issued guidelines focus, among other tests, on physicochemical characterisation of these molecules. The suggested analytical techniques, however, are only vaguely named in the specifications, leaving the final decision to the manufacturers. The present work focuses on the use of different combinations of analytical techniques with an aim to demonstrate similarity or dissimilarity between two or more samples. The selected instrumental techniques are characterised by their simplicity and are able to detect structural differences and microheterogeneity of the active ingredient in different samples, aggregation, degradation and post-translational modifications (PTMs). Seven studies were completed in total, each one to a different extent, and these included protein therapeutics such as insulin and monoclonal antibodies. The applied techniques served for primary (MS),[*] secondary (far-UV CD, FT-IR) and tertiary structure (near-UV CD, fluorescence) comparison of the examined samples. Particle size comparability and detection of aggregation was achieved with DLS, and higher-order structure comparison with 1D ^{1}H-NMR. Coupling of the techniques with temperature-dependent measurements enabled further comparison on the thermal stability of the samples and provided confidence in the observed (at room temperature) results. The acquired empirical experience pointed out the advantages and disadvantages of each technique compared to the rest of the techniques, possible solutions to the encountered challenges, and the cases that one technique can be used instead of another or as complementary to it. Finally, a potential SOP is suggested, advising on which

[*] MS: mass spectrometry, UV: ultraviolet, CD: circular dichroism, FT-IR: Fourier transform infrared, DLS: dynamic light scattering, NMR: nuclear magnetic resonance, SOP: standard operating procedure.

biophysical techniques should be used depending on the structure of the protein that is examined and its formulation.

Chapter 1
Introduction

The work reported in this thesis focuses on methods for structurally characterising proteins used as biopharmaceutical drugs. In order to provide a context for this work the focus of this chapter is on the products available on the market and how they have been reviewed in the United States and Europe. Protein structure and folding is overviewed and a brief description of the theoretical concepts of the applied analytical techniques for the structural characterisation of the proteins follows. Finally, the reader is introduced to the biotherapeutic proteins examined in this work and the case studies that are discussed in the next chapters.

1.1. An Introduction to Biopharmaceuticals and Biosimilars

Terminology

The term biopharmaceuticals or else biotechnological or biological products, was first introduced in the 1980s with the advent of recombinant DNA technology and refers to proteins, polypeptides and their derivatives that are produced from cell-culture expression systems. Unlike small molecules, which have a well-defined structure and in their vast majority can be chemically synthesised in large quantities and can be fully characterised, the manufacturing of biological products is highly complicated involving living cells, and therefore these products are required to be highly purified in order to meet safety standards.

In the late 1990s a new term, biosimilar, was introduced for the description of biological products that not only consist of the same active ingredient as the already-licensed one, but also are comparable with it in terms of structure, biological activity, immunogenicity, purity and quantity of the drug substance. In simple terms, a biosimilar product is the attempt of a company to "copy" the original biological product after the expiration of its patent, resulting in a highly similar product with the only differences allowed being those pertaining to their clinically inactive ingredients (excipients). Thorough testing of the biosimilar against a reference standard (usually the original product, also referred to as the innovator product) is required, including

structural and functional characterisation of the new product in order to confirm quality and consistency of the active substance and safety of its final formulation.

History

The advent of recombinant DNA (rDNA) technology in 1972 and monoclonal antibody technology[1] in 1975 permitted cloning and expression of proteins and peptides, and effectively led to the first genetically engineered insulin three years later. In 1982, the pharmaceutical company Eli Lilly & Co. launched the first FDA-approved (Food and Drug Administration) biopharmaceutical drug, with recombinant human insulin as the active ingredient and under the commercial name Humulin.[2] The first FDA-approved therapeutic monoclonal antibody followed only four years later from Janssen-Cilag (a subsidiary of the Johnson & Johnson pharmaceutical company) under the brand name Orthoclone.[3] This was a murine monoclonal antibody (muromonab), which acted as an immunosuppressant in cases of transplant rejection.

The advent of biomanufacturing and the following and ongoing achievements in biological processes such as cell culture and fermentation technologies, and engineering, have further accelerated the drug development process and made biological products cost-effective and safer than when extracted from tissue sources. An idea of the state of the rapidly changing and constantly increasing biopharmaceutical and biosimilar market in October 2016, and the "game of patents" is given in the next section.

1.1.1. Biopharmaceuticals and biosimilars in the market

Biopharmaceutical medicines have attracted the interest of the scientific community for more than three decades, with most of the big pharmaceutical companies expanding their Research and Development (R&D) strategies to biotechnology and biologics development. Biopharmaceuticals constitute currently the most expensive and advanced therapies for serious conditions such as cancer, autoimmune disorders (*e.g.* rheumatoid arthritis), blood-related diseases (*e.g.* anaemia) and neurodegenerative diseases (*e.g.* multiple sclerosis). Due to their high effectiveness and specificity, these drugs are safer and present fewer side effects than conventional therapies, with the most striking example that of targeted cancer therapies with monoclonal antibodies (mAbs) as opposed to traditional chemotherapy with cytotoxic agents.

Besides vaccines, the majority of proteins and peptides, oligonucleotides, blood-derived products and stem cells come under the umbrella of biologics. Examples are given in Table 1.1. They can be classified as follows:

1. Hematopoietic growth factors: a group of proteins that promote blood cell growth and bone marrow proliferation—*e.g.* erythropoietin, indicated for the treatment of anaemia (Epogen, Table 1.1).
2. Monoclonal Antibodies: antibodies produced by a single line of plasma cells, able to bind to specific antigens with various mechanisms of action—*e.g.* trastuzumab, indicated for the treatment of HER2-overexpressing breast and gastric cancers (Herceptin, Table 1.1).
3. Vaccines: biological preparations consisting of immunogenic agents (attenuated or inactivated microbes, inactivated toxins or segments of the pathogen) and providing active immunisation to a specific disease—*e.g.* pneumococcal conjugate vaccine with diphtheria protein, indicated for prevention of pneumonia and invasive disease caused by 13 *Streptococcus pneumoniae* strains (commercially available as Prevnar 13). While vaccines are better known for their use in a prophylactic manner, therapeutic vaccines also exist, and efforts are currently being made for the expansion of this area, with the main interest focused on targeted treatments for cancer, STDs (Sexual Transmitted Diseases, *e.g.* HIV and HPV) and hepatitis C. Sipuleucel-T (Provenge) is an example of a therapeutic vaccine with an indication for castrate resistant prostate cancer.
4. Thrombolytic agents: enzymes involved in the dissolution of blood clots—*e.g.* alteplase (tPA, tissue plasminogen activator), indicated for the treatment of ischemic stroke, myocardial infarction and pulmonary embolism (Activase).
5. Interferons: a group of signalling proteins produced by the cells of the immune system in response to various pathogens such as viruses or cancer cells—*e.g.* Interferon beta-1a, indicated for the treatment of multiple sclerosis (Avonex, Table 1.1).
6. Hormones: a class of signalling molecules of different chemical structures (amines, steroids, peptides, proteins) produced in glands and transported to specific organs where they express their regulatory activity—*e.g.* insulin, indicated for improvement of the glycemic control in diabetes mellitus (Lantus, Table 1.1).
7. Blood factors: a group of proteins that causes blood to clot in the coagulation cascade—*e.g.* coagulation factor VIIa, indicated for the treatment of bleeding

episodes and perioperative management in individuals with haemophilia (NovoSeven).

It is important to note though, that from the above-mentioned biological products FDA excludes hormones (*e.g.* insulin, human growth hormone, glucagon), which are therefore considered as drugs and are regulated by the *FDC Act* (*Federal Food, Drug and Cosmetic Act*), rather than by the *PHS Act (Public Health Service Act*) which includes the rest of the products. Biological products (including biosimilars) licensed by FDA under the *PHS Act* have currently reached the total number of 418 and are categorised in two separate lists, referred to as the "Purple Book".[4] In the CBER (Centre for Biologic Evaluation and Research) list,[5] 297 products are registered as vaccines and blood factors, while the remaining 121 biologics can be found in the CDER (Centre for Drug Evaluation and Research) list[6] and refer to therapeutic proteins and antibodies. Out of these 121 products, only 4 are biosimilars with the first (Zarxio) approved in June 2015 and the fourth (Amjevita, Table 1.2) in September 2016,[7] while 3 more have been withdrawn from the market.

The market in Europe is more complicated, as medicines can be granted authorisation following either the centralised or a national procedure. This means that EMA is responsible only for medicines that have been approved via the centralised authorisation procedure and not for all medicines that are available in Europe. National competent authorities of the respective EU Member States are responsible for medicines that have been approved via the national procedure.[8] EMA counts currently 259 biological products (holding centralised marketing authorisation) in total since 1995, with 66 of them registered as vaccines and blood factors, 163 as monoclonal antibodies and recombinant proteins (including insulin products), and 23 as biosimilar products. The first biosimilar was approved in 2006 (Omnitrope),[9] and 2 out of the 23 biosimilars have been withdrawn from the market.

The top-20 selling biopharmaceutical products for the year 2015, where the focus is on antibodies and recombinant proteins, are shown in Table 1.1. In order to be in accord with the FDA's CDER list of licensed biological products, vaccines (*e.g.* Prevnar), peptides (*e.g.* Copaxone), glucagon products (*e.g.* Victoza) and blood factors (*e.g.* NovoSeven) have been excluded from the table even if they held high positions in the sales ranking. Insulin products, however, have been included despite the fact that FDA does not consider these products as biologics since according to the European Medicines Agency (EMA), insulin products are still classified as biological medicinal

products and form a separate category from the rest of the biologics.[10] Monoclonal antibodies are now at the centre of attention for the development of both innovator products and biosimilars, as they bring considerable financial benefits to pharmaceutical companies besides their substantial benefit to public health. Only for the year 2015, 7 out of the top 10 drugs by sales were biologics, with 5 of them being monoclonal antibodies.[11]

Table 1.1: Top-20 biopharmaceutical products sales based on revenue for the year 2015.[12] Vaccines, peptides, glucagon products and blood factors have been excluded from the list, while insulin products are included.

No.	Product Name	Active Ingredient	Main Therapeutic Indication	Company Name	2015 Revenue in Billions (USD)
1	Humira	**Adalimumab**	Immunology	AbbVie	14.01
2	Enbrel	**Etanercept**	Immunology	Amgen/Pfizer Inc.	8.70
3	Remicade	**Infliximab**	Immunology	Johnson&Johnson/ Merck & Co	8.36
4	Rituxan (US)/ MabThera (EU)	**Rituximab**	Oncology	Roche	7.12
5	Lantus	**Insulin Glargine**	Diabetes	Sanofi	7.03
6	Avastin	**Bevacizumab**	Oncology	Roche	6.75
7	Herceptin	**Trastuzumab**	Oncology	Roche	6.60
8	Neulasta	**Pegfilgrastim**	Blood Related Disorders	Amgen	4.72
9	Lucentis	**Ranibizumab**	Ophthalmology	Novartis/Roche	3.60
10	NovoLog (US)/ NovoRapid (EU)	**Insulin aspart**	Diabetes	Novo Nordisk	3.11
11	Humalog	**Insulin lispro**	Diabetes	Eli Lilly	2.84
12	Levemir	**Insulin detemir**	Diabetes	Novo Nordisk	2.75
13	Avonex	**Interferon beta-1a**	Neuroscience/ Mental Health	Biogen	2.63
14	Soliris	**Eculizumab**	Blood Related Disorders	Alexion Pharmaceuticals	2.59
15	Stelara	**Ustekinumab**	Immunology	Johnson&Johnson	2.47
16	Aranesp	**Darbepoetin Alfa**	Blood Related Disorders	Amgen	1.95
17	Tysabri	**Natalizumab**	Neuroscience/ Mental Health	Biogen	1.89
18	Orencia	**Abatacept**	Immunology	Bristol-Myers Squibb	1.88
19	Epogen (US)	**Epoetin Alfa**	Blood Related Disorders	Amgen	1.86
20	Perjeta	**Pertuzumab**	Oncology	Roche	1.46

1.1.2. Patents

In contrast to the EMA, which has been granting approvals for biosimilars for years, FDA's legislation for such products was only recently established under the BPCI Act (Biologics Price Competition and Innovation Act of 2009) adopted by the Affordable

Care Act, which was enacted in 2010 (signed by President Obama). Though some rules are yet to be set, the new law clearly states that exclusivity rights for a new drug run for 12 years counting from the approval date. Patents, however, are issued by USPTO (United States Patent and Trademark Office) with a totally separate procedure, and at any time during the development of a drug or after approval. As a general rule, patents last for 20 years from the day that the application is filed. This regulation was specifically amended in order to avoid cases of drugs with overextended patents such as Amgen's Enbrel, which is protected until 2028, namely for 30 years since its approval date (Table 1.2). Despite the fact that Enbrel's patent was filed in 1995, it was only granted in 2011 due to delays in the Patent Office. This incident has so far worked in favour of Amgen, providing Enbrel 17 more years without competition counting from the date that the patent was issued rather than filed, since the specific patent followed the regulations of the old law, which was applicable when the patent was initially filed.[13,14]

However, this is not an obstacle for competitors, as they have the right to challenge the validity of patents. Sandoz's Erelzi, which is an Enbrel biosimilar (Table 1.2), gained FDA-approval in August 2016, before the official expiration of Enbrel's patents.[15] Launch in the market, though, can be delayed for at least six months from the approval date—according to US courts decisions—and up to a few years until the protecting patents are resolved, while at the same time the two companies have probably reached a confidential agreement.[16] Another case is that of Celltrion's Inflectra against its reference product J&J's Remicade (Table 1.2). Inflectra was the first monoclonal antibody biosimilar to get EMA-approval in 2013 and FDA-approval in April 2016, before Remicade's patent expiration. In the US market, one of Remicade's patents[17] was invalidated (August 2016) by a court decision, however, launching Inflectra would be a risk for Celltrion as another Remicade patent[18] case has yet to be resolved.[19,20]

The above cases are examples of how a biosimilar's introduction to the market can be delayed after its actual approval. The reasons may vary from court regulations that protect the reference product's (innovator's) rights to publication of new extended patents, covering *e.g.* formulations of the drug, manufacturing methods or additional indications and usages or even to hesitation of doctors and patients to switch to a therapy with a biosimilar product, especially when they have already seen improvement following a therapy with the original product. From the scientific point of view, the required tests and procedures might be enough to prove that a biosimilar is

highly similar to the innovator product, however, it is always the case that the products are not absolutely identical which affects public opinion as to their use.

Despite the vague aspects of the currently available guidelines for demonstrating biosimilarity (see Section 1.1.4), the risk of losses due to reimbursement of the originator company and the inevitable delays in biosimilar selling, many companies are specialised in developing biosimilars and invest hundreds of millions of dollars and 7 to 8 years in R&D. Emerging markets, such as that of India, are ideal for debuting a biosimilar as a low cost alternative, monitoring the product and collecting data with an ultimate goal of entering larger and more profitable markets, such that of the U.S.[21] For each reference product, tens of biosimilar medicines are in different stages of development or phases of clinical trials or even approved in some countries. The major advantage of introducing more than one product with the same active ingredient and indications in a market is the increased competition and therefore lower prices. Cheaper therapies with a reduced price of 30% on average,[22] but without discounts in safety and effectiveness compared to the original product, will be available and accessible to a wider public, and health care systems will benefit from the subsequent savings. Biosimilar developers claim that it is their duty to bring biosimilars, manufactured under high quality standards and GMPs (Good Manufacturing Practices), to patients as quickly as possible.

Table 1.2: Biopharmaceutical products from Table 1.1 for which biosimilars in Europe or the United States have been approved or withdrawals of approved products have been recorded. Approval dates and patent expiration dates are given, and the reference medicine where available.

Product Name	Active Ingredient	Date approved[1]	Patent Expiration Date[2]	Approved Biosimilars[1]	Reference Medicine[3]	Withdrawals (approved-withdrawn date)[1]
Humira	**Adalimumab**	2002 (US), 2003 (EU)	2016 (US), 2018 (EU)	Amjevita (US, 2016)	Humira	Trudexa (EU, 2003-2007)
Enbrel	**Etanercept**	1998 (US), 2000 (EU)	2028 (US), 2015 (EU)	Erelzi (US, 2016), Benepali (EU, 2016)	Enbrel	
Remicade	**Infliximab**	1998 (US), 1999 (EU)	2018 (US, invalidated in 2016), 2015 (EU)	Inflectra (US, 2016/ EU, 2013), Remsima (EU, 2013), Flixabi (EU, 2016)	Remicade	
Lantus	**Insulin Glargine**	2000 (US/EU)	2014 (US), 2014 (EU)	Basalgar (US, 2015), Abasalgar (EU, 2014)	Lantus	
Neulasta	**Pegfilgrastim**	2002 (US/EU)	2015 (US), 2017 (EU)			Neupopeg (EU, 2002-2008)
Aranesp	**Darbepoetin Alfa**	2001 (US/EU)	2024 (US), 2016 (EU)			Nespo (EU, 2001-2008)
Tysabri	**Natalizumab**	2004 (US), 2006 (EU)	2020 (US), 2023 (EU potentially)		in-house	Natalizumab Elan Pharma (EU, 2008 refused)
Epogen (US)	**Epoetin alfa**	1989 (US)	2013 (US)	Abseamed, Binocrit, Epoetin Alfa Hexal (EU, 2007)	Eprex/ Erypo (EU)	Eprex (US, 1999 discontinued)

[1] Information about approval dates, approved biosimilars and withdrawn products were extracted from the EMA and FDA drug databases by manual search of the active ingredient. The links for the search pages are following:
http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/landing/epar_search.jsp&mid=WC0b01ac058001d124
http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm?fuseaction=Search.Search_Drug_Name

[2] Patent expiration dates were taken from[23] and GABI Journal (Generics and Biosimilars Initiative Journal)[24] and updated by press publications (*e.g.* Remicade).

[3] Reference medicines were individually extracted from each product's EPAR (European Public Assessment Reports) from the EMA website.

1.1.3. Withdrawals, refusals and discontinuations. What went wrong?

Regarding cases where approved products were **withdrawn** from the market, three of the withdrawals mentioned in Table 1.2 (Trudexa, Nespo, Neupopeg) were due to commercial reasons, with Trudexa (adalimumab) not even making it to the market of

any member state in the European Economic Area (EEA). Namely, the marketing authorisation holder (MAH) has voluntarily decided to withdraw the under additional monitoring[*25] product for various reasons that presented it as unprofitable, and are explained below. Financial reasons is a quite common case of a withdrawal and even more common than "forced" withdrawal due to lack of efficacy or safety reasons, yet it does not attract public attention as much as the latter.

Trudexa, after several post-authorisation procedures, was indicated for the treatment of rheumatoid arthritis. EMA requested additional monitoring and further studies on the product due to uncertainties about the effects of long-term usage. Similarly, Nespo (darbepoetin alfa, Table 1.2) a medicine indicated for the treatment of anaemia, had to go through additional measures regarding concerns about its safety, which were related to unexplained excess mortality in cancer patients. In both cases, the provided data from the clinical trials were approved by EMA with a few modifications (*e.g.* in the wording of the package leaflet). For Trudexa, follow-up programmes were requested to continue as some of the data were lacking in sampling or in statistical significance, while Nespo was concluded to be favourable (with a positive benefit/risk profile) for the indicated treatments that were updated in the package leaflet to reflect a warning in the case of cancer patients.

Despite the fact that the data from the clinical trials for both of the above-mentioned products as well as for Neupopeg (pegfilgrastim, Table 1.2) were good enough to provide them with extensions of their marketing authorisations by EMA, the actual reasons why the companies (Abbott responsible for Trudexa and Dompé Biotec responsible for Neupopeg and Nespo) decided eventually to withdraw the medicines still remain unclear. In fact, all three of the here-mentioned withdrawn products (Trudexa, Neupopeg and Nespo) received their EMA-approval on very close dates if not on the exact same day as their respective (of the same active substances) and current "blockbusters" with the tradenames Humira, Neulasta and Aranesp (Table 1.2). What made the last three products "blockbusters" is the fact that these were also approved by FDA (as well as approved in Europe), while Trudexa, Neupopeg and Nespo received approval only in Europe. In all three cases (Trudexa/Humira, Neupopeg/Neulasta, Nespo/Aranesp), the post-marketing revisions of the products requested by the EMA were taking place in parallel for the eventually withdrawn product and its "blockbuster" counterpart, raising the same concerns and questions

[*] Medicines authorised in the European Union (EU) that are being monitored particularly closely by regulatory authorities.

about their safety.[26-31] It seems possible therefore, that the "European" product could not afford to compete with the one that had the extra advantage of being launched in the larger U.S. market, or to gather sufficient data from the clinical trials. Having estimated wrongly the market's needs, the companies which were responsible for the European versions of the medicines decided to withdraw their marketing authorisations, giving place to monopolies. Worth mentioning is also the fact that Dompé Biotec (responsible for Neupopeg and Nespo) and Amgen (responsible for Neulasta and Aranesp) became co-marketing partners in 2007, just one year before the respective withdrawals, while Abbvie (responsible for Humira) originated as a spin-off company of Abbott (responsible for Trudexa) in 2013.

A different case of withdrawal is that of Bexxar (tositumomab Iodine 131) in the U.S. market. The answer in that case is probably more complicated than a simple reason, and seems to be the outcome of several incidents and coincidences, which were extensively followed by Timmerman[32] and Prasad.[33] In summary, Bexxar was a novel radiotherapeutic antibody that received FDA-approval in 2003 and was indicated for the treatment of non-Hodgkin's lymphoma, where the disease was resistant to the standard therapy with rituximab and chemotherapy. Bexxar was never indicated as a first-line treatment, as its safety and efficacy after long-term usage had not been established. In the same way as for the above-mentioned cases, the requested post-authorisation studies (by FDA) got approved but the data acquired were not of statistical significance due to the limited number of the enrolled patients. This resulted from the fact that physicians were used to prescribing the already-tested rituximab and chemotherapy as a first choice therapy, rather than referring the patients to a radiation oncologist expert who could handle the new medicine. Even if Bexxar seemed to be a promising drug, the available data did not show a clear advantage over the pre-existing therapy (*e.g.* improvement in response rate or survival time), which in combination with the time-consuming clinical trials gave ground for the establishment of the standard therapy and the development of more advanced medicines such as ADCs (Antibody Drug Conjugates) which are mAbs attached to toxins rather than being radiolabeled. GlaxoSmithKline saw a major reduction in Bexxar's sales after 2006 and decided to withdraw the medicine in 2014.

Refused products as recommended by EMA, are those that have been refused marketing authorisation due to safety or efficacy reasons, and Natalizumab Elan Pharma (Table 1.2) is an example. In 2004, the company Elan Pharma submitted two separate applications for two products with the same active ingredient (natalizumab):

Natalizumab Elan Pharma was intended for the treatment of Crohn's disease when treatment with corticosteroids and immunosuppressant drugs was insufficient, and Tysabri was indicated for the treatment of multiple sclerosis. While the first was refused marketing authorisation (2008)[34] due to insufficient evidence of effectiveness and safety concerns about the risk of serious infections (such as brain infection), for the second the benefit to risk ratio was favourable for the suggested indication and it was approved by EMA (2006) even if it raised the same concerns about serious infections. Biogen purchased the MAH rights for Tysabri in 2013 with a variable per cent return to Elan, depending on the annual sales.[35]

According to FDA, products that have been withdrawn from the market for reasons other than safety or efficacy (voluntarily withdrawn according to EMA), approved products that have never been marketed or have been discontinued from marketing, and products that are for military use or for export only, are all named **discontinued** products. Eprex (epoetin alfa, Table 1.2) is one of the most important cases and has concerned not only the manufacturing company (J&J), but also academic research and national regulatory authorities that investigated different aspects of the drug's safety in the late 1990s to early 2000s. Eprex was the first medicine of the ESA class (erythropoiesis-stimulating agent) to receive EMA-approval in 1988 with the main indication being the treatment of anaemia due to chronic kidney disease (CKD). Epoetin alfa became available in the U.S. in 1989 under the tradenames Epogen (by Amgen, Table 1.2) and Procrit (by J&J), while Eprex got FDA-approval in 1999 but only for use outside the U.S.

In 1998, the European authorities requested a change in one of the excipients in Eprex's formulation, which was substitution of HSA (human serum albumin) with the stabiliser Tween 80, in the fear that HSA could potentially transmit Creutzfeldt-Jakob disease ("mad cow" disease) to Europe. Since then, the reported incidents to J&J of erythropoietin antibody-mediated PRCA (pure red cell aplasia) in patients with chronic renal failure (CRF) who had received Eprex subcutaneously, had significantly increased and seemed to be related mostly to Eprex than to other medicines of the same class.[36] In 2002 the unexpected excess mortality forced the company to impose intravenous administration to patients with CRF and ensure that sufficient attention is given to storage, handling and shipping under refrigerating conditions (the "cold chain").[37] In 2003, it was suspected that the reason for the increased immunogenicity of the product and PRCA was the release of leachates due to interaction of Tween 80 with the stoppers used in the pre-filled syringes, and as a result new rubber stoppers

(coated) were manufactured.[38] After this modification, the cases of PRCA were reduced back to the previously reported numbers of the well-tested formulation with HSA as a stabiliser.

Concerns about the ESA class, though, remained and FDA published a reassessment report of the risks of ESAs for the treatment of anaemia in cancer patients in 2007.[39] The data were insufficient to draw a clear conclusion on the harmful effects on survival in patients, and regular reassessments were decided to continue. In order to alert patients to the risks, FDA authorities proceeded in adding the "black box" warning in the package leaflet, which is the strongest label warning, underscoring the risk of blood clots, stroke, heart attack, tumour promotion and death. In addition, the prescription of ESAs was restricted under the risk evaluation and mitigation strategy (REMS) in 2010, which required that physicians should get additional training in order to assess high-risk users and negative side effects.

1.1.4. The importance of Quality Control (QC) in the development of biopharmaceuticals and the ICH guidelines

All here-mentioned cases of withdrawals remind us of the importance of clinical trials in establishing the safety and efficacy of a medicinal product. In order to assess any side effects, effects of long-term usage or the impact of a modification in the formulation or the administration method, sufficient data should be collected, and for that to happen a few years of post-marketing experience are necessary. Clinical trials (Phases I–III) and post-marketing studies (Phase IV) have the last word and are the most time-consuming "tests" for a medicine, with around 2% of the drugs that enter human clinical trials becoming officially approved medicines. Close pharmacovigilance control is of utmost significance in that stage, ensuring that the same product is used throughout the progress of a therapy, and no switching between products is taking place unless these have been evaluated as interchangeable. As a general rule, newly-authorised biological or biosimilar medicines are indicated for fewer or even different treatments than other authorised products with the same active ingredient or reference products, and therefore cannot substitute each other.

More specifically, Eprex's case is a striking example of how immunogenicity can increase with the slightest modification in the formulation of a product, the administration method or the used devices. Biological products are very complex molecules and therefore extremely sensitive to any kind of changes, from the expression and purification process to formulation and packaging, handling and storage conditions, and all of them can play a role in decreasing the stability of the final

product. As a consequence, the product becomes more susceptible to stress conditions, which could facilitate or accelerate denaturation of the protein or formation of aggregates, which in turn can increase immunogenicity and adverse effects. Considering these challenges and in order to develop a strategy for predicting and reducing immunogenicity of therapeutic proteins, especially when repeated administration is required, EMA's CHMP (committee for medicinal products for human use) issued guidelines (2005)[40] on how to characterise a proposed biological medicinal product and how to demonstrate similarity with the respective reference product. Since then, revised versions have been issued, as well as guidelines specific to each class of biosimilar products.

The current specifications as regulated by EMA (ICH Q6B, international council on harmonisation)[41] focus on determination of physicochemical properties, biological activity and purity of the biological product, naming a series of test procedures that should be fulfilled, but only vaguely mentioning the required analytical techniques. Data should be collected for different batches and in several stages of the development process (*e.g.* during preclinical and/or clinical studies, data for demonstration of manufacturing consistency, stability studies), aiming at confirming the quality of the product rather than fully characterising it. It is up to manufacturers, therefore, to decide on the most suitable techniques that would provide scientific data sufficient to support a positive safety and efficacy profile. Subtle differences between the biosimilar and the reference product, between biosimilars from different manufacturers or between different manufacturing processes are very likely to occur, however, the level of variance can be acceptable if cutting-edge analytical, manufacturing, clinical and regulatory procedures are adopted. Even though experience from the above-mentioned examples has shown that some adverse effects might be apparent only during post-authorisation studies and after long-term usage of the medicine, quality control and more specifically physicochemical characterisation, which is the aim of the present study, can provide a certain amount of confidence and filter out the majority of these cases, that otherwise would have had much more serious effects.

Analytical techniques that are reported in this work for the physicochemical characterisation of purified proteins or formulated products, are able to detect structural differences and microheterogeneity of the active ingredient in different samples (batch-to-batch comparison or compared to the reference product), aggregation, degradation (*e.g.* deamidation, oxidation) and post-translational

modifications (*e.g.* glycosylation profile), all of which have been previously blamed in the literature for increasing immunogenicity.[42] More details on the techniques used and the cases that were investigated in this study are given below (Sections 1.3 and 1.4), following a brief introduction on protein structure and folding (Section 1.2).

1.2. PROTEIN STRUCTURE AND FOLDING

A protein's native structure usually refers to the acquired conformation of the protein following correct folding, which is a fundamental requirement for its function and stability. The three-dimensional atomic structure of a protein is primarily determined by its amino acid sequence. In order for a protein to fold into its native structure, essentially the thermodynamically stable structure, all types of intermolecular forces—covalent and non-covalent—play a role, with the most important of these being hydrophobic interactions and hydrogen-bonds. In the early stages of protein folding, non-polar amino acids are driven away from water and towards the interior of the protein (hydrophobic collapse), while more polar residues are exposed to the aqueous environment of the solvent. The side chains of the amino acids are the reason that the folding of one protein differs from another and Van der Waals interactions have a large contribution during this stage, resulting in tight packing of the protein. In the next stage, hydrogen-bonds among backbone amides and carbonyl groups are formed, resulting in more compact helical or sheet configurations, which stabilise the secondary structure of the protein. The formation of disulfide bridges results in further stability of the secondary structure, while the tertiary structure depends mainly on interactions among the side chains and finalises the three-dimensional structure of the protein.

Since folding is a transition from disorder to order, the occurring conformational changes aim at the most energetically minimised structure of the protein. The fact, however, that proteins have funnel-shaped energy landscapes, namely many high- and few low-energy states, renders protein folding a very complicated process to predict or simulate. A funnel describes the conformational heterogeneity among the non-native states of a protein, namely the multiple folding routes, which also depend on the used conditions (such as solvent, pH and glycosylation), as well as the chain entropy of a protein, which increases as the structure becomes less stable.[43] The balance between the chain entropy and the forces of folding leads to denaturation of the protein, which is favoured under progressively altered conditions such as high temperatures (thermal denaturation) due to the presence of many high- as opposed to fewer low-energy states. During thermal denaturation (*i.e.*

in protein stability studies), the melting temperature (T_m) of the protein is defined as the midpoint of the unfolding transition, where the free energy of folding becomes zero, while above this temperature the unfolded protein "occupies" a high-energy state. Thermal denaturation is often irreversible as it can lead to non-native folds or trapped intermediate structures along the folding pathway (local minima in the energy funnel), and also due to aggregation of the heat-unfolded polypeptide (see Sections 3.5.1 and 4.7.2). In many cases misfolding of the protein at high temperature can lead to formation of β-sheet rich fibrils that resemble the structure of natural amyloid (see Section 2.6).

1.3. USED TECHNIQUES AND COMPLEMENTARITY

The main goal of this work was to establish the extent to which different techniques could be used to evaluate rapidly whether or not different samples had the same structure. The selected instrumental techniques are characterised by their simplicity, relatively low time consumption, ease of sample preparation, and because they provide different and complementary information.

Mass Spectrometry (MS) is an established technique and most commonly used for structural confirmation and characterisation of biopharmaceuticals, covering a wide range of information necessary for lot release purposes. Giving answers to fundamental questions, such as solving the primary structure, looking into post-translational modifications (carbohydrate structure and position, sulfhydryl groups and disulfide bridges *etc.*) and degradation products (*e.g.* oxidation, deamidation), mass spectrometry is the technique of choice for companies offering protein characterisation services. Although very small quantities of sample are sufficient for MS analysis, as a technique it is more time consuming than the rest (see below) and usually requires additional steps in sample preparation (denaturation, reduction, digestion—as discussed in Chapter 3), which are moreover irreversibly destructive for the protein.

Thus, the use of complementary techniques is essential in order to determine the spectroscopic profile and examine the higher-order structure of a product. Circular dichroism (CD) spectroscopy is the simplest way to estimate the secondary structure of proteins in solution (using far-UV data from 260 nm to 190 nm or lower) and to acquire a tertiary structure spectroscopic pattern (using near-UV data). Dynamic light scattering (DLS) is probably the simplest way to estimate protein size and to detect any aggregates or multimers in the intact sample. Nuclear magnetic resonance (NMR) spectroscopy of proteins is commonly used to give atomic resolution structures;

however, in this comparability work easily accessible 1D ^{1}H NMR was chosen for its ability to probe variations in atomic environments and folding of the protein. Fluorescence spectroscopy can quickly show any differences in the environment of the fluorescent residues. Fourier-transform infrared spectroscopy (FT-IR) is mainly used for comparing secondary structures in a qualitative manner and a tentative trial to quantitate the results was attempted (discussed in Sections 2.6 and 4.6.3). Ultraviolet-visible spectroscopy (UV-Vis) is usually used for determination of the extinction coefficient; here, it was used for estimating the concentration of a protein based on its known extinction coefficient. Finally in the case of insulin (Chapter 2), molecular dynamics (MD) simulations complemented these instrumental techniques, by giving indications of how structures can change with temperature. As with NMR, MD was simply used as an indicator of the effect of heating the protein and not in a sophisticated way.

1.3.1. Mass spectrometry (MS)

The basic principle of a mass spectrometer is as follows: the **sample inlet** transfers the sample to the mass spectrometer and directs it to the **ion source**. In the ion source, the sample molecules are transformed into gas phase ions, which are accelerated by an electromagnetic field and separated by the **mass analyser** depending on their mass-to-charge ratio (*m/z*). The signal of the ions that reach the **detector** is recorded and processed by the **data system**, usually a PC, which produces the **mass spectrum** (number of ions detected as a function of their *m/z* ratio). Often, the mass spectrometer is connected to a chromatograph (gas or liquid) in order to facilitate separation of complex mixtures and, subsequently, determination of the individual mass spectrum of each eluted substance.

Mass spectrometry became a very popular technique for the study of high molecular weight biomolecules in the mid-1980s, with the advent of matrix assisted laser desorption ionisation (MALDI), which was invented by Hillenkamp *et al.*,[44,45] and electrospray ionisation (ESI),[46] which was developed by John Fenn. Tanaka, who further developed MALDI, and Fenn were jointly awarded one half of the Nobel Prize in Chemistry in 2002.[47] A description of ESI follows, as this method was used in the present work for the ionisation of the produced peptides. ESI is a soft ionisation method, capable of generating multiple charged ions of intact protein molecules[48] with very little fragmentation. In brief, the sample (analyte dissolved in volatile buffer/solvent) is sprayed through a high-voltage potential capillary into the heated ionisation chamber (close to atmospheric pressure). The details of the mechanism for

going from solution to ions are still debated, but for large molecules such as proteins, it can be thought of as follows. The resulting charged droplets are subjected to a drying gas (usually nitrogen), which progressively evaporates the solvent (Figure 1.1, left). As the amount of solvent decreases, the charge density of each drop increases and the repulsive forces lead to further dissociation into smaller droplets. Complete evaporation of the solvent leaves sample ions in the gas phase, bearing a single or multiple charges, depending on molecular size and other parameters. The charges are often provided by protons, as proteins are usually analysed with the ESI-MS analyser in the positive ion mode, with protonation of different sites of the protein or the peptide being the reason of the multiple charges. As mentioned above, the sample ions are accelerated by an electric field and directed into the mass analyser in order to get separated and detected.

Various mass analysers are available, however, here only the principal operations of the time-of-flight (TOF) and the quadrupole analysers are briefly explained, since these types were used in the present work. The TOF analyser measures the time that ions of different masses need until they hit the detector. Effectively this is assuming that all ions are produced at the same time and have the same kinetic energy. Therefore light ions will arrive first and heavier ions will follow. In order to increase the resolution of this analyser and correct for the slightly different kinetic energies, an ion reflector is often used, namely an electric field which acts as a mirror by reversing the route of the ions towards the ion source and allowing for longer flight times (Figure 1.1, top right).

The quadrupole mass analyser works as a “mass filter” using a dynamic electric field, generated by applying a direct-current voltage (DC) and a radiofrequency (RF) to four solid rods with parallel arrangement to the direction of the ion beam. Depending on the ratio RF:DC, ions acquire an oscillation in this electrostatic field and those with a specific *m/z* range travel through the device following a stable trajectory and reach the detector, while the rest of the ions undergo an unstable oscillation and are lost prior to detection (Figure 1.1, bottom right).

A triple quadrupole tandem mass analyser consists of three quadrupole devices arranged in a row (TQ-S), facilitating MS/MS experiments to be carried out. The first quadrupole transmits ions of certain *m/z* value (precursor or parent ions), determined by the operator. The second device is used as a collision cell, where fragmentation of the parent ions occurs due to their increased internal energy, typically yielding b- and y- ions (described below). All ions (parent and fragment) are then transmitted into the

third quadrupole, which is scanned to obtain a fragment ion spectrum. Triple quadrupole mass spectrometers are necessary in applications that require quantitation of both small molecules and biological analytes (see Section 3.3.2).

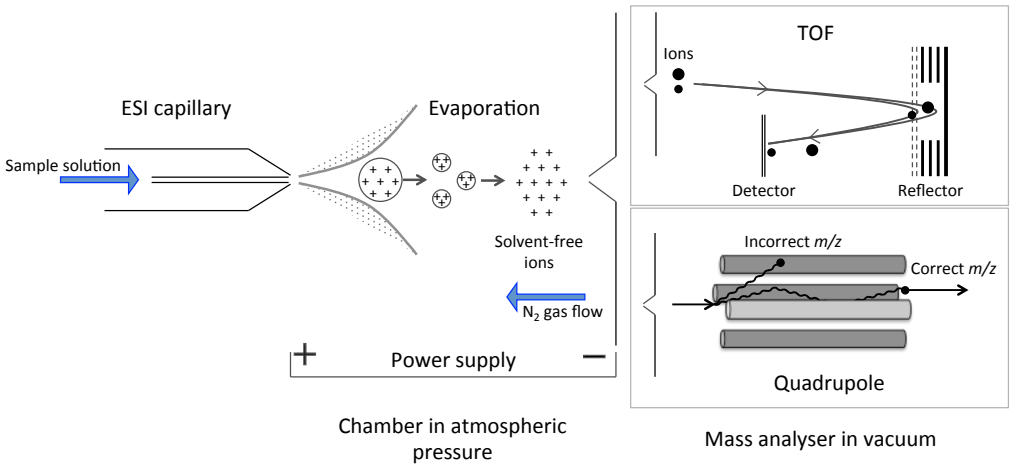


Figure 1.1: Schematic representation of the basic principle of a mass spectrometer with ESI sample introduction. Left: the sample is sprayed through a capillary into a heated chamber. Once the solvent evaporates from the droplets, the ions are accelerated by an electromagnetic field and directed to the mass analyser for separation. Top right: a TOF analyser is shown with a reflector. Bottom right: a quadrupole analyser is depicted.

A Q-TOF mass spectrometer is a hybrid instrument, consisting of a quadrupole and a time-of-flight analyser. It is a very common configuration as it provides data of better quality for the time and expense of instrumentation. The front-end (quadrupole) is used for selecting the precursor ions, the following collision cell (using only RF) generates the fragments, which are then analysed by the TOF section with a reflector, offering high-resolution. The detector consists of a counter that produces a current proportional to the number of ions that strike it. Since only a few ions make it to the detector, electron multiplier circuits are normally used, which essentially produce two electrons every time an ion strikes the electron multiplier surface. A series of collisions occur all the way until the end of the detector, resulting in a signal amplification of 2^n (n is number of collisions). The recorder, eventually, produces the mass spectrum, which is stored in a computer.

Fragment ions of biological molecules, and more specifically in proteins, usually result from cleavage of a single covalent bond along the backbone. Depending on the type of the bond cleaved and whether the fragment ion contains the N- or C-terminal portion of the peptide, the ion is classified as *a-, b-, c- or x-, y-, z*-type respectively (Figure 1.2). Fragment ion peaks are often found in the mass spectrum of the parent ion of the peptide, due to low energy fragmentation, and aid determination

of the amino acid sequence of the peptide. In tandem mass spectrometry ion dissociation is carried out following isolation of the ion of interest, which is subjected to collisional activation (as described above for TQ-S), allowing for the parent-to-fragment ion transition to be followed.

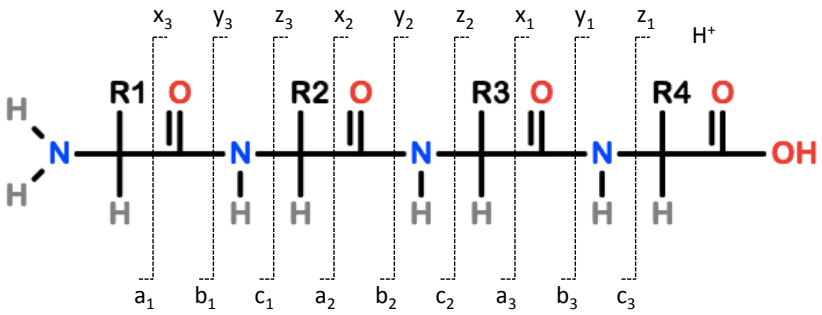


Figure 1.2: Fragment ions generated in proteins. Depending on the bond of the backbone that is cleaved and whether the resulting fragment ion contains the N- or the C-terminus of the peptide, the ions are classified as *a-, b-, c- or x-, y-, z*-type respectively. The peptide structure was sketched in chemspider.com (http://www.chemspider.com/StructureSearch.aspx)

Biopharmaceutical products are often glycosylated proteins (*e.g.* monoclonal antibodies), and glycosylation is one of the major sources of heterogeneity in such molecules, as well as potential immunogenicity of the product (explained in more detail in Chapter 3). Determination of the primary structure of the protein (amino acid sequence) and identification of the attached glycans (structure and position) are therefore among the most fundamental requirements in the ICH guidelines. For protein sequencing, the bottom-up approach is usually used, during which the protein is enzymatically cleaved, the resulting peptides are separated by liquid chromatography (LC) and subsequently analysed by MS and/or MS/MS methods, allowing for great sequence coverage in a single experiment. At the same time, structural analysis of the carbohydrate moieties of a glycoprotein, and localisation of the glycosylation sites is simplified in electron-based ionisation techniques, which most often cleave the peptide backbone leaving the carbohydrate chains intact.[49]

1.3.2. Absorption spectroscopy

Ultraviolet-Visible spectroscopy (UV-Vis)

Electronic excitations

Light is a rapidly oscillating electromagnetic field that can be described either as a wave or a photon. Molecules respond to the interaction with light by absorbing certain amounts of electromagnetic radiation and allowing the rest of the light (that is not

absorbed) to be transmitted through them, or by scattering light (discussed in Section 1.3.4). The absorbed energy, in turn, results in transitions of the molecule's electrons to various energy levels of an excited state, while the subsequent return of the electrons to the ground state results in the release of this energy in different forms (heat generation, fluorescence, *etc.*). These phenomena, namely the absorbance of light, the emitted energy as well as the light that is scattered, are exploited by different spectroscopic techniques, such as UV-Vis absorbance (ultraviolet-visible), CD (circular dichroism), FT-IR (Fourier transform infrared), NMR (nuclear magnetic resonance), fluorescence and DLS (dynamic light scattering), providing information on the analysed molecules and more specifically on the conformation and folding of biological macromolecules such as proteins.

To a reasonable approximation the different levels of energy result from the formation of different chemical bonds. The paired electrons of a bond occupy the low energy bonding orbitals (σ or π), any unshared electrons occupy the nonbonding orbital (n), while the antibonding orbitals (π* and σ*), at higher energy, are empty. As a result, molecules usually exist in a low energy state (ground state) with their electrons located in bonding and nonbonding orbitals, while incident light of suitable frequency can cause transition of the molecule to a temporary state of higher energy (excited state) as its electrons ascend to the antibonding orbitals (Figure 1.3). The relation between the absorbed energy and the frequency (or wavelength) of the incident light is described by Planck's equation:

$$E = h\nu = \frac{hc_0}{\lambda} \quad (1.1)$$

where h is the Planck constant (6.626×10^{-34} J s^{-1}), ν is the frequency (s^{-1}), c_0 is the speed of light (3×10^8 m s^{-1}) and λ is the wavelength (m).

The most frequent transitions are those that take place from the highest occupied molecular orbital (HOMO) to the lowest unoccupied molecular orbital (LUMO). In the case of protein molecules, the observed (accessible) absorbance results from the π to π* (peptide backbone and aromatic side chains) and n to π* (peptide backbone) transitions (Figure 1.3), as discussed below in the UV-Vis and CD sections. It is important to mention that the UV absorption occurs over a wide range of wavelengths, as each electronic state consists of a large number of vibrational and rotational modes, which are energetically very close to each other. As a result, exciting electrons from each of these states to the respective higher energy states will result in

slightly varying amounts of absorbed energy (at varying wavelengths). Thus, the produced UV spectrum will consist of a broad peak that represents the sum of all the arising transitions and the different events occurring at each molecule, providing the average for the total population. Similarly, the emission spectrum, as the molecules return to the ground state, will also result in a broad band.

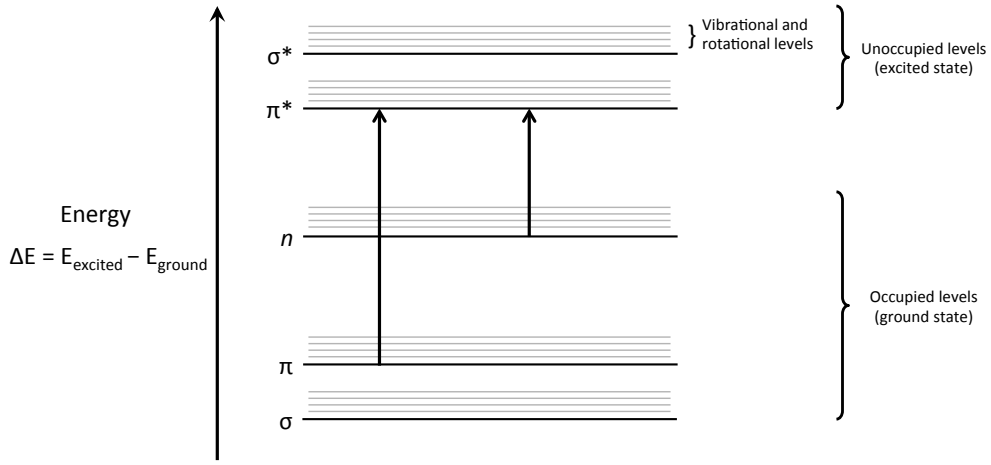


Figure 1.3: Energy levels of the ground and excited states of a molecule. The bonding (σ, π), nonbonding (n) and antibonding (π*, σ*) orbitals are shown with their vibrational and rotational levels (only a few of them are shown for clarity). The arrows between the orbitals (from π to π* and from n to π*) refer to the most commonly observed transitions in protein molecules.

As described above, biological macromolecules, and more specifically proteins, present absorbance in the UV region as their electrons ascend to higher energy levels. The bonds or groups of a protein that are able to absorb UV light (chromophores) at specific wavelengths are mainly: the carbonyl group of the peptide bond (at ~190 nm and 210–230 nm), aromatic amino acid side chains (250–280 nm) and disulfide bonds (<260 nm) (described in the CD section).[50] The typical absorbance spectrum records the amount of light absorbed by the analysed sample in a selected range of wavelengths, while for protein quantitation, the most commonly used wavelength is at 280 nm as contribution of the used buffer excipients is minimised in that region compared to lower wavelengths.

The broad peak at 280 nm results mostly from the contributions of tryptophan (~280 nm) and tyrosine (~276 nm) residues and disulfide bonds (<260 nm) and it is widely used for determining the concentration of proteins[51] by applying the Beer-Lambert law:

$$A = log_{10}(\frac{I}{I_0}) = \varepsilon cL \quad (1.2)$$

where A is the measured absorbance, I_0 is the intensity of the incident light, I is the intensity of the light that is transmitted through the sample, ε is the extinction coefficient (M^{-1} cm^{-1}), c is the molar concentration of the protein (M) and L: is the distance that the light travels through the material (path length, cm).

The concentration of a protein can therefore be determined, provided that an accurate extinction coefficient for the specific protein is obtained. The extinction coefficient is the absorbance of a 1 M solution of the contributing chromophores at the wavelength maximum (*e.g.* 280 nm) and can be either theoretically calculated or experimentally measured.[52] The calculated or measured extinction coefficient is usually an approximation as this can vary depending on the environment of the chromophores in the protein, namely interactions with nearby amino acids and exposure to the solvent, while the wavelength maximum is dependent on the aromatic amino acid composition and also the individual environments of the chromophores. Finally, the determined concentration might slightly diverge from the actual concentration when larger protein aggregates or traces of nucleic acids are present.

Circular Dichroism (CD)

Circular dichroism is based on the interaction of an asymmetric (chiral) molecule with circularly polarised light and measures the difference in absorption of left- and right-circularly polarised light by that chiral molecule, resulting in a CD spectrum with positive and negative peaks. Biological macromolecules, such as proteins, consist of a series of chiral units (backbone and amino acid side chains), which further contribute to the formation of asymmetric three-dimensional structures such as helices and pleated sheets (secondary structure) depending on the hydrogen bonding patterns. The fact that these units and structures interact to a different degree with left- and right-circularly polarised light gives rise to characteristic CD spectra, representative of the dominant secondary structure (far-UV region) or the environment of the aromatic amino acid side chains and disulfilde bonds (near-UV).

A circularly polarised light beam results when two linearly polarised light beams incident along the same direction but polarised perpendicular to each other, of equal magnitude and out of phase by π/2, are combined (Figure 1.4). Therefore, the resulting light beam propagates by following a helical path (either right- or left-handed) of a constant magnitude. Two circularly polarised beams, one with a clockwise (right-handed) and one with an anticlockwise (left-handed) direction, are shined onto an isotropic (*e.g.* solution) sample during a CD experiment. If the sample

is symmetric (with no chiral centres) the difference of the two lights will be zero, as both lights will be absorbed to a similar extent. A chiral molecule, however, will respond differently to each of the light beams, absorbing one more than the other and therefore resulting in a positive or negative signal. As a consequence, the remaining light for each light beam (light that was not absorbed) will be of different magnitude when the beams will be exiting the sample, resulting in an elliptically polarised light when subtracted from each other. This elliptical light is recorded by the CD detector and translated to a CD spectrum over a wide range of wavelengths.

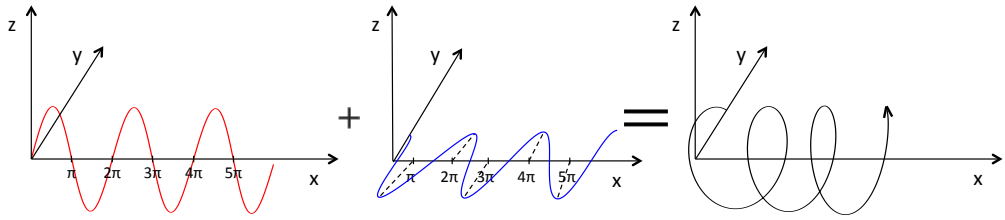


Figure 1.4: Left and centre: two linearly polarised beams (in red and blue) of equal magnitudes and propagating in perpendicular planes with a phase difference of π/2, combine to give a circularly polarised beam (right, in black). The circularly polarised beam shown here is left-handed, following an anticlockwise helix around x axis.

As with UV-Vis absorbance the prerequisite for such a phenomenon is that the initial incident light beams are of suitable energy (and therefore wavelength), able to cause redistribution of the molecule's electrons. However, in the case of CD, the electrons follow a helical transition (the displacement of electronic charge has a circular and a linear component in the direction of the applied field) due to combination of the electric and magnetic fields of the radiation as opposed to the linear rearrangement that takes place in UV-Vis as a result of interaction predominantly with the electric field.[53] Protein CD spectra are divided into two regions, the far-UV (180–250 nm) and the near-UV (250–350 nm), which provide different information on the conformation of a protein depending on the interacting chromophores. The CD signal in the far-UV region results from the occurring electronic transitions in the peptide backbone, namely the π to π* (at ~190 nm)[54] and n to π* (210–230 nm)[55] transitions, which give rise to characteristic CD spectra for different secondary structures (*e.g.* helices, sheets) as they are affected by the respective hydrogen bonding patterns in each structure.[56-58] Transitions of the side chains of the aromatic amino acids can also present CD signals in the far-UV region; however, the magnitude is smaller compared with the backbone signals except for some unusual peptides such as gramicidin which has a high percentage of tryptophan residues.[59] Electronic transitions of the aromatic

amino acid side chains and disulfide bonds are significant in the near-UV region.[50,56] More specifically, π to π* transitions occur at: ~280 nm for tryptophans,[60] ~276 nm for tyrosines,[61,62] 250–270 nm for phenylalanines[63] and <260 nm for disulfide bonds.[64] The resulting CD spectrum in the near-UV region is affected by the number of aromatic amino acids in the protein sequence and the local environments of the individual aromatic amino acids (such as interactions with nearby amino acids and polarity). Thus, the near-UV CD spectrum is distinct for each protein and characteristic of the folding and its tertiary structure.[65,66]

Similarly to any spectroscopic technique, a CD spectrum is representative of the average of the molecular population and the contributing structures or events that take place in solution. CD finds wide application in studying conformational changes and unfolding (or folding) of proteins under varying conditions such as temperature, pH and protein-ligand interactions, providing valuable qualitative comparative studies on the secondary (far-UV CD) and tertiary (near-UV CD) structure of the protein. CD data are usually reported in Δε units (mean residue ellipticity), which facilitates comparison of the CD spectra of proteins with largely different molecular weights.

$$\Delta\varepsilon = \frac{\theta \times \mathrm{MRW}}{\mathrm{C} \times \mathrm{P} \times 32980} \qquad (1.3)$$

where $\Delta\varepsilon$ is molar differential extinction coefficient (M^{-1} cm^{-1}), θ is machine units (mdeg), C is mass concentration (mg mL^{-1}), P is path length (cm), and MRW (mg $mmol^{-1}$) is mean residue weight, which for the peptide bond is calculated as: MRW = molecular mass of protein (Da) / (number of amino acids − 1).[67]

Several programs are used for the analysis of far-UV CD spectra and are available through various web sites such as DichroWeb[68,69] and CDPro[70] or as stand-alone versions, *e.g.* SSNN (secondary structure neural network).[71,72] All of them use datasets of CD spectra of proteins, whose secondary structures have been solved by X-Ray crystallography or NMR experiments, in order to find the contribution of each secondary structure component (with a well-defined CD spectrum) to the measured CD spectrum and translate it into quantitative percentages. Three of these programs were used in this work for the estimation of the secondary structure components of different proteins, and on some occasions at different temperatures. SSNN was used in all cases (Chapters 2–4), SELCON3 (self-consistent)[73,74] was used in Chapters 2 and 4, and CDSSTR[75,76] was used in Chapter 4. All three are further discussed in Chapter 4.

Fourier Transform Infrared spectroscopy (FT-IR)

Infrared (IR) spectroscopy measures the wavelength and intensity of the infrared light that is absorbed by a sample, able to cause transitions in the vibrational state of a molecule. IR and therefore excitation of the molecular vibrations to higher energy levels requires less energy than the electronic transitions that occur in UV. The commonly used IR unit is the wavenumber (cm^{-1}), which is the reciprocal of the wavelength in cm, and subsequently directly proportional to the absorbed energy. Therefore equation 1.1 becomes:

$$E = hc_0\tilde{\nu} \qquad (1.4)$$

where $\tilde{\nu}$ is the wavenumber (m^{-1}) and the rest of the parameters are as described for equation 1.1. Incident light of suitable energy, in a wavenumber range of 4000–400 cm^{-1}, increases the vibrational motions of the chemical bonds (*e.g.* stretching or bending) in a molecule. However, only bonds that have a dipole moment that varies over time are able to absorb IR radiation, whereas symmetric bonds (*e.g.* H_2) do not interact with infrared light. The strength of the chemical bond, the contributing atoms, and the change in the dipole moment determine the intensity and vibrational frequency of an IR band, therefore IR spectra give information on the characteristic bonds present in a molecule.

For proteins, the most prominent (and repetitive) bond is the amide bond of the peptide backbone, which presents a number of bands in different regions of the FT-IR spectrum. The three most important amide bands are:

- the amide I, which is mainly associated with the C=O stretching vibration and to a lesser extent with C-N stretching. It depends on the conformation of the protein and the H-bonding, and appears at around 1650 cm^{-1}.
- The amide II band results from in-plane N-H bending and C-N stretching vibrations and is also sensitive to the protein conformation. It appears at around 1550 cm^{-1}.
- The amide III bands are more complicated, resulting from C-N stretching and C-N-H in-plane bending modes. They appear at around 1200–1350 cm^{-1}.

Out of amide I, II and III, the amide I band is the most sensitive to conformational changes and gives information on the secondary structure of a protein,

as different secondary structure components have characteristic band frequencies in this region (1600–1700 cm^{-1}).[77] Antiparallel β-sheet structures, for example, present an average frequency of 1633 cm^{-1} and a second frequency at around 1684 cm^{-1}, α-helix appears at ~1654 cm^{-1} on average, turn structure has an absorption near 1672 cm^{-1} and unordered structure presents a band close to the α-helix frequency, at ~1654 cm^{-1}.[78] Several methods (*e.g.* second derivative) are often used for resolving the amide I band into its individual components, allowing for a quantitative analysis of the secondary structure of a protein. These are briefly discussed in Chapter 2 (Section 2.6).

In this work, attenuated total reflectance Fourier transform IR spectroscopy (ATR FT-IR) was used. In Fourier transform IR (FT-IR) spectroscopy, an interferogram is produced (time-domain spectrum) by moving mirrors and converted to the frequency-domain spectrum by a mathematical operation called Fourier transform. When a number of interferograms are accumulated and summed, the signal-to-noise ratio increases. In attenuated total reflectance (ATR) FT-IR spectrometry the sample is placed in contact with an internal reflection element (IRE) of a higher refractive index, *e.g.* a ZnSe crystal. The radiation is directed into the crystal and towards the sample and the method is actually based on the reflection of the infrared beam at the sample-crystal interface and in the IRE medium, while the number of reflections varies from single to multiple, depending on the length of the crystal. A prerequisite for the reflection of the beam is that the angle of incidence is equal or greater than the critical angle (θ_c), namely the angle of the incident light (θ_1) that would theoretically result in a refraction angle (θ_2) of 90°. The critical angle can be calculated from Snell's law, assuming that $\theta_2 = 90°$:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \rightarrow \theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \qquad (1.5)$$

where n_1 is the refractive index of the crystal and n_2 refractive index of the sample. Practically, as the angle of incidence increases up to the θ_c value, the refracted beam will approach 90° and will be reflected when θ_c is reached (Figure 1.5, left). The light beam is reflected with an angle equal to the angle of incidence and directed to the detector.

ATR is an advantageous method as a variety of samples can be used (*e.g.* in solution, absorbed on a surface or in solid state) with no sample preparation. However, two effects related to the technique, the penetration depth and the anomalous

dispersion, result in a somewhat different spectrum compared to a transmission spectrum, where the light passes through the sample. As the light beam hits the sample-crystal interface, the electrical field of the incident photon extends perpendicularly from the surface of the IRE into the sample, resulting in an evanescent wave with a frequency similar to that of the incident light, but with an amplitude that decays exponentially with distance from the surface of the crystal (Figure 1.5, right). The **depth of penetration** (d_p) is the effective distance from the surface, at which the evanescent wave can interact effectively with the sample, and is linearly dependent on the wavelength. This effect results in distorted relative intensities of the peaks of an ATR spectrum, as peaks at high wavenumbers will have smaller intensities than peaks at the low-wavenumber end. The effect of the **anomalous dispersion** results from the variation of the refractive index across an absorption band, leading to a shift of the absorption bands to lower wavenumbers.[79,80]

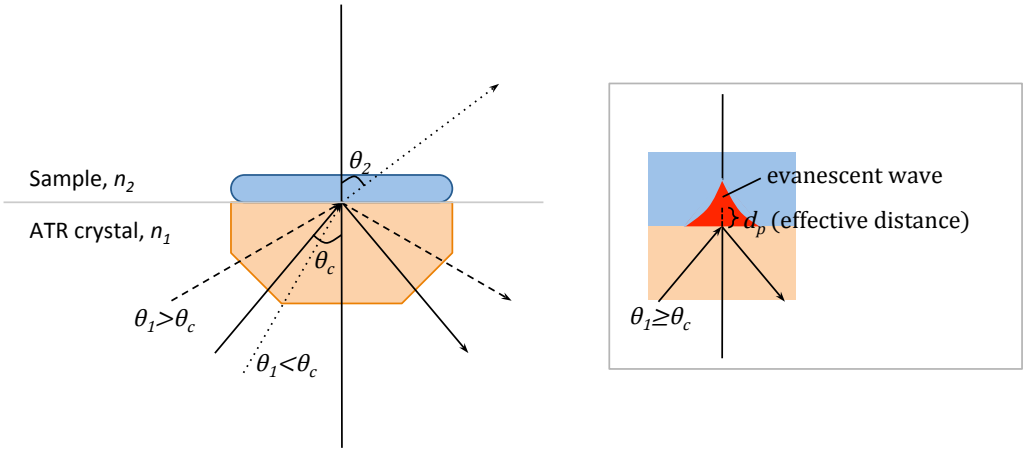


Figure 1.5: Crystal (in orange) and sample (in blue) arrangement in an ATR FT-IR experiment. Left: incident light beams of different angles are shown, resulting in refraction (when $\theta_1<\theta_c$, dotted lines) or reflection (when $\theta_1\geq\theta_c$, solid and dashed lines) of the light beam. Right: Zoom on the sample-crystal interface at the point where the incident light beam hits. When $\theta_1\geq\theta_c$ the light is reflected, while its electric field creates an evanescent wave (red) that expands perpendicularly from the sample-crystal interface towards the sample, and decays exponentially with distance. The penetration depth is the distance from the surface, at which the evanescent wave can still interact with the sample.

Among its numerous applications, FT-IR spectroscopy is one of the preferred techniques for secondary structure characterisation and conformational analysis of therapeutic proteins.[81,82] Studies on protein folding and stability are quite common, allowing for a qualitative comparison by monitoring changes in the secondary structure of the molecule induced by various factors (*e.g.* solution conditions),[83-85] while deconvolution of the amide I band region to its individual components can provide quantitative results on the secondary structure of a protein.[77,86]

Nuclear Magnetic Resonance spectroscopy (NMR)

In NMR spectroscopy, the use of radiofrequency (RF) radiation combined with a static magnetic field cause transitions of the nuclear spins of an atom. Only atomic nuclei with odd mass, odd atomic number or both present non-zero nuclear spin, with the allowed spin states being quantised and determined by the nuclear spin quantum number I. For each nucleus, I is a physical constant and the allowed spin states (magnetic quantum number, m) can take values from $-I$ to $+I$ in the following way: $-I$, $-I+1$, …, $I-1$, I. For example, ^{1}H (I = ½) can occupy either of the two spin states m = –½ and m = +½. When a static magnetic field is applied, the allowed spin states become energetically distinguishable and transitions between these states can be captured by NMR. The frequency of such transitions depends on the strength of the magnetic field and the observed nucleus and is determined by the following equation:

$$2\pi v = \omega = \gamma B_0 \qquad (1.6)$$

where v is the transition frequency in s^{-1} (Hz), ω is the transition frequency also referred to as Larmor frequency (radians s^{-1}), γ is the gyromagnetic ratio (a constant that depends on the identity of the nucleus) and B_0 is the strength of the magnetic field (Tesla). Due to its high natural abundance and favourable gyromagnetic ratio, the ^{1}H isotope is the nucleus of choice in many NMR experiments.

In order to generate the correct amount of energy for the nuclear spin transitions, **RF pulses** are used, which are defined by their frequency (called the carrier frequency, ν_{RF}), power, phase and duration. When a pulse of a specific ν_{RF} is applied, a range of frequencies are also excited on both sides of the carrier frequency. The range of these frequencies is determined by the pulse duration, with short pulses leading to a wider range of frequencies and longer pulses resulting in more focused excitation (narrow spectral region). The pulse power measures the energy intake by the sample and the NMR components (*e.g.* NMR probe) per unit time, with low and high power pulses being combined in NMR experiments, but high power pulses of increased length being avoided in order to protect the sample and the probe from the destructive consequences. Since a pulse has the nature of a wave, the pulse phase is defined by the amplitude of the wave at the beginning of the pulse (t = 0). For a sine wave, for example, with maximum amplitude at t = 0, the phase shift is $\pi/2$.

Similarly to all the absorption spectroscopic techniques mentioned above, the energy of the applied RF pulse is absorbed by nuclear spins in the sample, resulting in

a transition from a lower to a higher energy state. In the presence of an external magnetic field, a slight majority of the hydrogen nuclei orient in respect to the applied magnetic field (B_0), adopting the lower energy +½ spin state—as the –½ spin state requires higher energy since it is opposed to the magnetic field—while at the same time they precess around the applied magnetic field at their transition frequency, called the Larmor frequency (Figure 1.6). The sum of magnetic moments for the nuclei with spin greater than 0 results in a net magnetisation along the magnetic field direction (+z axis), which tilts away from the axis as soon as an RF pulse is applied, resulting in a net magnetisation that precesses around z, in the xy plane.[87] The precession of the net magnetisation around the z axis induces an electric current in the coil of the NMR probe, which is actually what is detected as a signal in a pulsed NMR experiment. The decay of this signal over time (which is a combination of the two relaxation events explained below), as the magnetisation returns to equilibrium (along the z axis), is called the free-induction decay signal (FID). Since the FID signal is recorded in the time-domain, Fourier transformation is applied, allowing for conversion into the frequency-domain spectrum.[88]

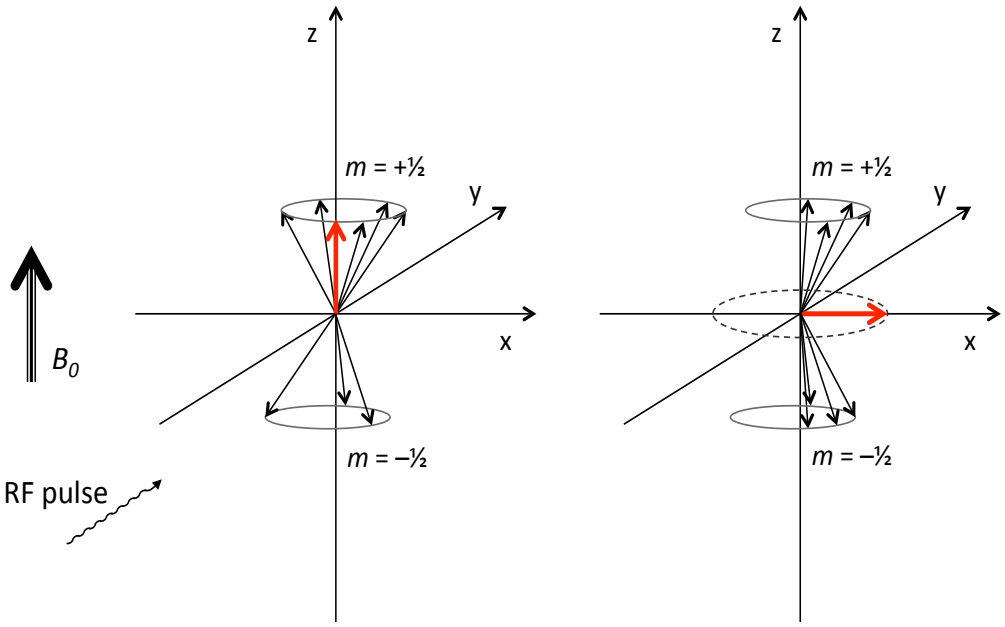


Figure 1.6: Left: spins precessing at their Larmor frequency at both spin states (m = –½ and m = +½). The sum of vectors from the individual nuclei results in a net magnetisation (in red) in alignment with the external magnetic field (B_0). Right: the applied radiofrequency pulse results in phasing the spins (coherence), which subsequently tilts the net magnetisation along the x axis, while at the same time some nuclei change their spin state from low to high energy.

Since the different spin states of a nucleus are very close energetically, NMR uses accordingly small energies to cause transitions (much smaller than what is

required in electronic and vibrational excitations). The absorbed RF energy results in a transition of the spins from the low to the high energy spin state until the two states are equally populated, at which point the transition is saturated and no more energy can be absorbed until the spins return to those populations of the upper and lower energy levels as predicted by the Boltzmann distribution. The z-magnetisation, therefore, is proportional to the population difference between the two spin states, and the constant of proportion determines the absolute size of the observed NMR signal. Acquisition of successive signals is delayed by the time that is required for the NMR signal to return to its equilibrium state, namely for re-establishment of the normal population distribution of the two spin states. This process is called **spin-lattice relaxation** and is accompanied by the release of energy to the local environment. The corresponding time is called spin-lattice relaxation time (T_1).

A second process related to relaxation is the decay of the transverse magnetisation of the spins, called **spin-spin relaxation**. As mentioned above, the applied RF pulse results in a net magnetisation (*e.g.* on the x axis) that precesses around z axis and in the xy plane at the Larmor frequency (Figure 1.6).[87] This orientation of the magnetisation is disturbed, however, by local spin-spin interactions (small magnetic fields from neighbouring nuclei), which lead to fluctuations in the local magnetic field and a slightly different precession frequency for each proton. As a result, the protons come out of phase (loss of coherence), which is what causes decay of the observed signal, namely decrease of the net magnetisation in the xy plane. The time required for the decay of the coherence-induced signal (to a certain amplitude) is called spin-spin relaxation time (T_2) and is an important factor in multi-pulse experiments of biological macromolecules, affecting the line widths of the spectrum in an inverse manner. The slow tumbling of larger macromolecules in solution leads to faster relaxation of transverse magnetisation (short T_2), due to enhanced spin-spin interactions, and therefore to broad line widths. This problem is often solved by recording NMR spectra at elevated temperatures, which result in an increase of the overall molecular tumbling rate. Coherences that last longer result in narrower line widths, while they can be also transferred to other spins via coupling, which is what gives rise to multidimensional NMR experiments.

The **chemical shift** is the variation in the resonant frequencies of spins, resulting from differences in their chemical environments. Such differences are caused by the electron density that surrounds the observed nuclear spin, and by the interactions of these electrons with the applied magnetic field. Subsequently, these

interactions result in fluctuations in the local magnetic field of the nucleus, and can either increase the net magnetic field sensed by the nucleus, moving the resonance to higher frequencies (deshielding), or decrease it, leading to lower frequencies (shielding).[89] The ratio of the difference in resonance frequency of the observed nucleus relative to a reference one (usually tetramethylsilane, TMS) is what is reported as chemical shift δ in ppm (parts per milion). The chemical shift of ^{1}H is largely affected by environmental factors (*e.g.* temperature, pH, hydrogen bonding, ring current effects), to the point that an ^{1}H NMR spectrum can reveal whether a protein is folded, presenting dispersed chemical shifts, or unfolded, where resonances for the same chemical type of nuclei (*e.g.* backbone NH protons or nitrogens) are observed in a relatively narrow range, due to the decreased local interactions and hence decreased differences in the chemical environment of these nuclei.[90]

As mentioned above, the generated coherences on one spin during an NMR experiment are usually long enough to be transferred to a coupled spin, therefore allowing the correlation between coupled spins to be probed. The coherence transfer can be achieved via **J-coupling** (or dipolar coupling, explained below), which is the result of the interactions between nuclei coupled through bonds and mediated by the electrons of the bond. The energies of the spinning electrons note a slight variation, depending on the spin state of the nuclei with which they are interacting, with these changes in energy sensed in turn by nuclear spins of the same bonded system. This effect results in somewhat different energy levels of the nuclear spin states, which depend on the coupled nucleus, and consequently in the separation of the signals with a similar "distance" between the signals for both nuclei, called coupling constant.

Multi-dimensional NMR experiments exploit the effect of coupled nuclei that are one or even a few bonds apart. Such experiments (*e.g.* 2D-, 3D-NMR) are nowadays used for the sequential assignments of proteins, with the structure of a macromolecule having been determined for the first time by Wüthrich in 1985.[91] A relatively simple and very commonly used homonuclear 2D-experiment is the TOCSY (total correlation spectroscopy),[92,93] from which spin systems of different amino acids can be identified and correlated. TOCSY is based on the Hartmann-Hahn effect, which takes place when a long RF pulse of low power is applied. This pulse results in spin-locking, namely freezing of the free precession of individual coherences, which consequently remain in phase while at the same time energy transfer occurs between coupled coherences. At the end of spin-locking, the detected coherences will result in

signal amplitudes that will be regulated in accordance with the frequencies of all the spins that participated in this energy transfer.

Heteronuclear 2D-NMR experiments are also extensively used, exploiting **polarisation transfer** from the ^{1}H nucleus to another NMR-active nucleus coupled to it, such as ^{13}C or ^{15}N. A heteronuclear experiment is the HSQC (heteronuclear single-quantum coherence),[94] for example, with the resulting spectrum showing correlation peaks between the two observed nuclei, when the nuclei in this experiment are connected to each other via one or more chemical bond(s). One more homonuclear experiment, but exploiting a different mechanism of energy transfer between nuclear spins than J-coupling, is the NOESY (nuclear Overhauser effect spectroscopy).[95] This experiment is based on **dipolar relaxation**, which is an interaction between spins that are close in space (typically with a distance <5 Å) and do not necessarily belong to the same spin system or even to the same molecule. The dipolar mechanism is based on the fact that each spin has a magnetic moment, resulting in a magnetic field which can interact with other spins. This interaction is a property of a pair of nuclei, and its size depends strongly on the distance between the two spins and the direction of the vector joining the two nuclei. As a molecule tumbles in solution, the direction of this vector changes and so does the magnetic field. Changes in the distance between the two nuclei also result in a change in the magnetic field, while in general changes in orientation are responsible for relaxation. In dipolar relaxation, therefore, energy can be transferred between the lattice (molecular motion) and the spins. As a result, dipole-dipole interaction turns molecular motion into an oscillating magnetic field, which can cause transitions of the spins.[96] In NOESY the cross-relaxation is responsible for exchange of magnetisation between different spins, and this energy exchange takes place during the **mixing time**.[*97] Thus, a NOESY spectrum presents cross-peaks of two spins (typically referring to ^{1}H-^{1}H correlations) that are experiencing mutual cross-relaxation and hence are close in space.

NMR is an extremely useful tool for the investigation of the structure and dynamics of biomolecules. Among many applications and advanced experiments, sequential assignment, which can be achieved with the acquisition of selected 2D- and 3D-NMR experiments, forms the basis of three-dimensional structure determination of

[*] Mixing time (or period) is one of the key parameters in 2D NMR spectroscopy as it allows for the transfer of magnetisation or, in other words, transfer of coherence among spins (through one of the two mechanisms described above) prior to acquisition. Depending on the type of the experiment, the mixing period consists of one or more pulses and delays.

biomolecules. However, one of the most significant limitations of the technique is the size of the protein molecule, with proteins larger than 30 kDa rendering challenging the acquisition of 2D (or of higher dimension) NMR spectra. Large proteins can be tackled, nowadays, with the use of sophisticated experiments such as methyl-TROSY (transverse relaxation optimised spectroscopy)[98] or FROSTY-MAS (**F**reezing **R**otational diffusion **O**f protein **S**olutions at low **T**emperature and high viscosit**Y** **M**agic-**A**ngle-**S**pinning spectroscopy),[99,100] in combination with optimised labeling methods and/or highly deuterated proteins. Although the use of 2D NMR experiments has been reported in the literature for the comparison of biosimilars with the corresponding reference products,[101] such complicated NMR experiments and protein labeling (^{15}N, ^{13}C) are very labour- and time-consuming processes, and therefore are not yet fully established for the purpose of quality control of formulated therapeutic proteins. Compared with 2D experiments, which are more complicated to set and interpret, 1D ^{1}H-NMR experiments provide a simple fingerprint of the analysed sample. 1D ^{1}H-NMR experiments were mostly used in this work in order to provide information on the qualitative comparison of the spectra of different samples of a peptide (insulin, Chapter 2) and a monoclonal antibody (Chapter 4).

1.3.3. Emission spectroscopy

Fluorescence

As mentioned above (UV-Vis in Section 1.3.2), absorption spectroscopy measures the energy that is absorbed by a molecule during its excitation from the ground state to an excited state. Fluorescence, however, is an emission spectroscopic method and is based on the measurement of the reverse process, namely the energy that is radiated as the electron returns to the ground state. The emitted light, in such cases, is of lower energy than the light that was absorbed during excitation as an amount of energy is lost by non-radiative processes. As a result, the emission spectrum is observed at longer wavelengths. According to Jablonski,[102] loss of energy is the result of internal conversion processes, which occur prior to emission. In other words, the molecules initially relax to the lowest vibrational level of the excited state and then return to a higher vibrational level of the ground state by the emission of photons, which is what causes the fluorescence phenomenon. The shift of the emitted light to longer wavelengths, compared to the incident light, was first observed by Stokes (Stokes shift).[103]

Proteins have intrinsic fluorescence, due to the fluorescence of the aromatic amino acids tryptophan, tyrosine and phenylalanine, with tryptophan presenting the most dominant fluorescence signal in most of the cases.[104-106] Similarly to the absorption spectrum, the released energy (emission spectrum) is also affected by interactions between the molecular dipole and the surrounding environment, namely interactions with nearby amino acids and solvent exposure. The high sensitivity of tryptophan, in particular, makes fluorescence valuable in monitoring conformational changes in proteins, as its fluorescence signal can be almost entirely isolated by the contribution of other aromatic amino acids when the absorption wavelength is set at 295 nm. The emission maximum of tryptophan usually provides information on protein folding, as the hydrophobic environment of a folded protein limits solvent exposure, resulting in a shift towards lower wavelengths (blue-shift) compared to the maximum of free tryptophan in water (353 nm). Consequently, unfolding and increased solvent exposure shifts the maximum towards longer wavelengths (red-shift).

Intensity of the fluorescence signal also depends on the local environment. Non-radiative processes are inhibited in the hydrophobic environment of a folded protein, therefore generally resulting in increased emitted energy (higher intensity) compared to the unfolded protein,[107] while in other cases Trp fluorescence is quenched by nearby amino acids in the folded protein, therefore showing lower intensity compared to the unfolded protein.[108] Fluorescence intensity is also affected by high protein concentrations, where an inner filter effect takes place due to the high density of the protein, resulting in decreased intensity compared to a diluted solution. Finally, the width of the band reflects the number of Trp residues and the heterogeneity of their environments in the folded protein.[50]

Due to its sensitivity as a technique and among other more advanced applications, fluorescence is largely used for studies on protein conformation, stability and folding or unfolding kinetics, providing information on the average environment of tertiary structures in the analysed sample. In combination with other spectroscopic techniques (CD, FT-IR, NMR *etc.*) it offers valuable information on protein structure, as discussed in Chapter 4.

1.3.4. Scattering spectroscopy

Dynamic Light Scattering (DLS)

Light scattering occurs when an incident photon induces a dipole moment in the electron cloud of a molecule, which in turn leads to the energy being radiated or

scattered in all directions.[50,109] Scattering can follow either **Rayleigh** or **Mie** theory, depending on the size of the molecule.[110] When the particle is much smaller than the wavelength of the incident light ($<\lambda/10$), it is assumed that all sides of the interacting particle are equally subject to the electric field, therefore resulting in isotropically scattered waves, of equal intensities in all directions (Rayleigh theory). For larger particles, the interactions become more complex with the scattered waves being distorted from the front (the side facing the incident beam) to the back of the particle (Mie theory).

DLS (dynamic light scattering) measures the intensity fluctuations of the scattered light from a collection of particles in a sample as a function of time. These fluctuations result from the motion of the particles in solution (Brownian motion) and intermolecular interactions.[111] As a result, a fast moving particle will cause rapid intensity fluctuations as opposed to a slow moving particle. The velocities of the molecules are related to their size, with small particles moving faster than larger particles, so particle size affects the DLS signals. In order to extract information about the particle size from DLS, the instrument produces an autocorrelation function of the fluctuating light intensity with respect to time. This function is then fit using two different algorithms in order to provide the weighted mean hydrodynamic size of the particles (Z-average) and polydispersity index (PDI)[112] or the particle size distribution.[113] The size distribution shows the relative intensity of light scattered by particles of different size classes (intensity distribution). The polydispersity index is related to the width of the size distribution and is a way of describing the uniformity of a sample. A PDI value <0.1 indicates an almost ideal, monodisperse sample with particles of one only size class present (a single exponential decay is fitted to the autocorrelation curve), while values >0.4 originate from broad distributions, indicating a highly polydisperse sample (the sum of multiple exponential decays is used).

In the case that there is more than one peak present (more than one particle population), the generated intensity distribution will be largely affected by the larger particle size, as large particles scatter more light than smaller particles and the intensity of the scattered light is proportional to the sixth power of the particle's diameter (Rayleigh approximation). Mie theory is then applied in order to convert the intensity distribution to a number distribution, assuming that the particles are of spherical shape.[114] The number distribution shows the number of particles in each population and is useful for comparative purposes in the case of polydisperse samples, but should not be used for reporting the absolute size of the particles.

As mentioned above, the detected fluctuations in the intensity of the light result from the random movement of the particles in solution and their bombardment by the solvent molecules (Brownian motion). The velocity of this motion is defined by the diffusion coefficient (D) of the particle in a particular solvent environment, thus depending on the size of the particle, the surface structure, concentration and the ions present. The diffusion coefficient is calculated from the autocorrelation function, since it is proportional to the lifetime of the exponential decay, and the hydrodynamic radius (R_H) is calculated using the Stokes-Einstein[115] equation:

$$D = \frac{kT}{6\pi\eta R_H} \qquad (1.7)$$

where k is the Boltzmann constant, T is the absolute temperature and η is the viscosity of the solution. The temperature should be accurately set, as any changes in the temperature will result in variation in the viscosity of the sample and therefore in miscalculation of the particle size. The final value of the particle size reported by the DLS instrument is the hydrodynamic diameter, which refers to the diameter of a spherical particle (including the hydration layer) that diffuses with the same speed as the measured molecule.

DLS finds use in the determination of the particle size of macromolecules, in aggregation studies and in forced degradation studies (*e.g.* as a function of temperature), where conformational changes and unfolding of the protein are followed via the detected changes in its diffusion. DLS is a nondestructive technique and particularly useful in probing the effects of solution conditions and formulation excipients on the stability of a therapeutic protein, while the polydispersity index gives information on the homogeneity of a sample. An increased PDI in the case that a single population is detected suggests either the presence of trace amounts of larger species or the presence of differently shaped molecules (*e.g.* due to unfolding). The sensitivity of DLS in detecting very small amounts of larger particles makes it a valuable tool for qualitative comparison of samples of biotherapeutic molecules, particularly indicating early onset of aggregation as discussed in Chapter 2 (Section 2.4).

1.4. Case Studies and Challenges

The above-described techniques were applied in different combinations on several protein samples and various case studies, which are reported in detail in Chapters 2, 3 and 4. The examined protein therapeutics include: insulin, trastuzumab, somatropin,

etanercept and a monoclonal antibody. The common goal in all studies was to demonstrate similarity or dissimilarity between two or more samples in particular to spot differences in their structure and detect aggregates or degradation products that if significant could potentially result in loss of activity or increased immunogenicity.

All samples, except insulin, were provided already formulated by the manufacturing company for comparability assessment, and were tested under normal (room temperature) conditions and in some cases under the induced stress of increased temperatures. Comparison of the samples at elevated temperatures allowed the detection of structural (and consequently stability) differences that were not obvious or might have been overlooked if the comparison had only taken place at room temperature.

The use of already formulated products in these tests is usually preferred to reconstitution for various reasons such as ensuring consistency in concentration and solvent environment, maintaining stability of the active substance, minimising sample preparation and human error. Furthermore, the formulation of a product with appropriate and compatible excipients usually requires specific knowledge[116-118] and several trials[119] in order to achieve solubility, reduce intermolecular interactions such as adsorption and aggregation, prevent microbe growth, avoid changes in pH, ionic strength and viscosity, situations that if not properly addressed could affect the product's stability and consequently safety and efficacy.

All examined cases were targeted at one common purpose: finding the combination of techniques that are able to demonstrate (or not) similarity between samples with confidence; what makes each case distinct is the reason why the examined samples might vary. In **Chapter 2** for example, a comparison of insulin samples under different conditions is discussed. Exploiting insulin's ability to exist in various oligomerisation states depending on the environmental conditions, a study was conducted that focused on investigating how stability and secondary structure are affected by oligomerisation state. **Chapter 3** comprises a detailed comparison between two mAbs (trastuzumab) and their respective ADCs that came from different conjugation processes. The main question was how and to what extent an alternative conjugation method could affect the structure of the antibody and the number of the conjugates attached to it. **Chapter 4** consists of various cases that include: structure determination of somatropin and batch-to-batch and/or biosimilar to reference product comparability studies (etanercept, mAb in phase III of clinical trials). The additional objective of these short studies of Chapter 4 was to suggest solutions to challenges,

technique-related or associated with the formulation or sample preparation, someone may come across when performing such analyses.

References

1. R. J. Y. Ho and M. Gibaldi, in *Biotechnology and Biopharmaceuticals*, John Wiley & Sons, Inc., 2004, DOI: 10.1002/0471704210.ch1, pp. 1-9.
2. L. K. Altman, A new insulin given approval for use in U.S., http://www.nytimes.com/1982/10/30/us/a-new-insulin-given-approval-for-use-in-us.html, (accessed 4 Oct, 2016).
3. The Pink Sheet, Ortho gets first FDA okay for therapeutic monoclonal: Orthoclone OKT3 approved for renal transplants, target population of 4,000 per year, https://pink.pharmamedtechbi.com/PS010346/ORTHO-GETS-FIRST-FDA-OKAY-FOR-THERAPEUTIC-MONOCLONAL-ORTHOCLONE-OKT3-APPROVED-FOR-RENAL-TRANSPLANTS-TARGET-POPULATION-OF-4000-PER-YEAR, (accessed 4 Oct, 2016).
4. U.S. Food and Drug Administration, Purple Book: Lists of Licensed Biological Products with Reference Product Exclusivity and Biosimilarity or interchangeability Evaluations, http://www.fda.gov/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/ApprovalApplications/TherapeuticBiologicApplications/Biosimilars/ucm411418.htm, (accessed 5 Oct, 2016).
5. U.S. Food and Drug Administration, Center for Biologic Evaluation and Resarch, http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/ApprovalApplications/TherapeuticBiologicApplications/Biosimilars/UCM412398.pdf, (accessed 5 Oct, 2016).
6. U.S. Food and Drug Administration, Center for Drug Evaluation and Resarch, http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/ApprovalApplications/TherapeuticBiologicApplications/Biosimilars/UCM439049.pdf, (accessed 5 Oct, 2016).
7. U.S. Food and Drug Administration, FDA approves Amjevita, a biosimilar to Humira, http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm522243.htm, (accessed 5 Oct, 2016).
8. European Medicines Agency, Authorisation of medicines, http://www.ema.europa.eu/ema/index.jsp?curl=pages/about_us/general/general_content_000109.jsp&mid=WC0b01ac0580028a47, (accessed 6 Oct, 2016).
9. European Medicines Agency, European public assessment reports, http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/landing/epar_search.jsp&mid=WC0b01ac058001d124&searchTab=searchByAuthType&keyword=olaratumab&searchType=inn&alreadyLoaded=true&status=Authorised&status=Withdrawn&status=Suspended&status=Refused&jsenabled=false&searchGenericType=biosimilars&orderBy=authDate&pageNo=1, (accessed 5 Oct, 2016).

10. European Medicines Agency, Guideline on non-clinical and clinical development of similar biological medicinal products containing recombinant human insulin and insulin analogs, http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2015/03/WC500184161.pdf, (accessed 5 Oct, 2016).
11. PharmaCompass, Top drugs by sales revenue in 2015: Who sold the biggest blockbuster drugs?, http://www.pharmacompass.com/radio-compass-blog/top-drugs-by-sales-revenue-in-2015-who-sold-the-biggest-blockbuster-drugs, (accessed 10 Oct, 2016).
12. PharmaCompass, Product Sales Data from Annual Reports of Major pharmaceutical Companies// 2015, http://www.pharmacompass.com/pharma-data/product-sales-data-from-annual-reports-of-major-pharmaceutical-companies-2015, (accessed 7 Oct, 2016).
13. A. Pollack, Patent for Amgen drug may undercut health care plan, http://www.nytimes.com/2011/11/23/business/amgens-new-enbrel-patent-may-undercut-health-care-plan.html, (accessed 7 Oct, 2016).
14. A. Pollack, Makers of Humira and Enbrel using new drug patents to delay generic versions, http://www.nytimes.com/2016/07/16/business/makers-of-humira-and-enbrel-using-new-drug-patents-to-delay-generic-versions.html, (accessed 7 Oct, 2016).
15. Z. Brennan, FDA approves third biosimilar in US, first for Amgen's blockbuster Enbrel, http://www.raps.org/Regulatory-Focus/News/2016/08/30/25739/FDA-Approves-Third-Biosimilar-in-US-First-for-Amgens-Blockbuster-Enbrel/, (accessed 7 Oct, 2016).
16. J. Miller, Novartis bid to sell new biosimilar crimped by U.S. court battles, http://www.reuters.com/article/us-novartis-fda-biosimilars-idUSKCN11619F, (accessed 7 Oct, 2016).
17. J. Le, J. Vilcek, P. Dadonna, J. Ghrayeb, D. Knight and S. A. Siegel, Anti-TNFa antibodies and assays employing anti-TNFa antibodies, https://www.google.com/patents/US6284471, (accessed 7 Oct, 2016).
18. D. Epstein, R. Monsell, J. Horwitz, S. Lenk, S. Ozturk and C. Marsh, Chemically defined media compositions, https://www.google.com/patents/US7598083, (accessed 7 Oct, 2016).
19. D. Stanton, Pfizer plans for October Remicade biosimilar launch, boosted by IP ruling, http://www.biopharma-reporter.com/Markets-Regulations/Pfizer-plans-for-October-Inflectra-launch-boosted-by-IP-ruling, (accessed 7 Oct, 2016).
20. D. Stanton, J&J: 'We can't stop Pfizer launching Remicade biosimilar in the US', http://www.biopharma-reporter.com/Markets-Regulations/J-J-We-can-t-stop-Pfizer-launching-Remicade-biosimilar-in-the-US, (accessed 7 Oct, 2016).
21. E. A. Blackstone and P. F. Joseph, *American Health & Drug Benefits*, 2013, **6**, 469-478.
22. B. Yu, *American Journal of Managed Care*, 2016, **22**, 378.
23. G. Walsh, *Nature Biotechnology*, 2014, **32**, 992-1000.
24. M. Derbyshire, *GaBI Journal*, 2015, **4**, 178-179.

25. European Medicines Agency, List of medicines under additional monitoring, http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000366.jsp, (accessed 20 Jun, 2017).
26. European Medicines Agency, Trudexa - Procedural steps taken and scientific information after the authorisation, http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Procedural_steps_taken_and_scientific_information_after_authorisation/human/000482/WC500058942.pdf, (accessed 13 Oct, 2016).
27. European Medicines Agency, Neupopeg - Procedural steps taken and scientific information after the authorisation. Changes made after 01/02/2004, http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Procedural_steps_taken_and_scientific_information_after_authorisation/human/000422/WC500055681.pdf, (accessed 13 Oct, 2016).
28. European Medicines Agency, Nespo - Procedural steps taken and scientific information after the authorisation. Changes made after 01/09/2003, http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Procedural_steps_taken_and_scientific_information_after_authorisation/human/000333/WC500055307.pdf, (accessed 13 Oct, 2016).
29. European Medicines Agency, Neulasta - Procedural steps taken and scientific information after the authorisation, http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Procedural_steps_taken_and_scientific_information_after_authorisation/human/000420/WC500025944.pdf, (accessed 13 Oct, 2016).
30. European Medicines Agency, Aranesp - Procedural steps taken and scientific information after the authorisation, http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Procedural_steps_taken_and_scientific_information_after_authorisation/human/000332/WC500026145.pdf, (accessed 13 Oct, 2016).
31. European Medicines Agency, Humira - Procedural steps taken and scientific information after the authorisation, http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Procedural_steps_taken_and_scientific_information_after_authorisation/human/000481/WC500050869.pdf, (accessed 13 Oct, 2016).
32. L. Timmerman, Why good drugs sometimes fail: The Bexxar story, http://www.xconomy.com/national/2013/08/26/why-good-drugs-sometimes-fail-in-the-market-the-bexxar-story/, (accessed 13 Oct, 2016).
33. V. Prasad, *JAMA Internal Medicine*, 2014, **174**, 1887-1888.
34. European Medicines Agency, Questions and answers on recommendation for the refusal of the marketing authorisation for Natalizumab Elan Pharma, http://www.ema.europa.eu/docs/en_GB/document_library/Summary_of_opinion_-_Initial_authorisation/human/000624/WC500070716.pdf, (accessed 14 Oct, 2016).
35. M. Herper, In wise move Biogen spends $3.25 billion for rest of Tysabri. Now what happens to Elan?, http://www.forbes.com/sites/matthewherper/2013/02/06/biogen-to-buy-elans-tysabri-rights-for-3-25-billion/ - 283680c9bca2, (accessed 14 Oct, 2016).

36. G. Kelley, *BioPharm International*, 2002, **15**, 44.
37. Johnson & Johnson, Intravenous administration required when using Eprex/Erypo (epoetin alfa) in chronic renal failure patients, http://www.investor.jnj.com/secfiling.cfm?filingid=200406-02-11&cik=, (accessed 18 Oct, 2016).
38. K. Boven, J. Knight, F. Bader, J. Rossert, K.-U. Eckardt and N. Casadevall, *Nephrology Dialysis Transplantation*, 2005, **20**, iii33-iii40.
39. U.S. Food and Drug Administration, Continuing reassessment of the risks of Erythropoiesis-Stimulating Agents (ESAs) administered for the treatment of anemia associated with cancer chemotherapy, http://www.fda.gov/ohrms/dockets/ac/07/briefing/2007-4301b2-02-FDA.pdf, (accessed 18 Oct, 2016).
40. European Medicines Agency, Guideline on similar biological medicinal products, http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003517.pdf, (accessed 19 Oct, 2016).
41. European Medicines Agency, ICH Topic Q 6 B Specifications: Test Procedures and Acceptance Criteria for Biotechnological/Biological Products http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002824.pdf, (accessed 1 Jun, 2017).
42. S. Sethu, K. Govindappa, M. Alhaidari, M. Pirmohamed, K. Park and J. Sathish, *Archivum Immunologiae et Therapiae Experimentalis*, 2012, **60**, 331-344.
43. K. A. Dill, S. B. Ozkan, M. S. Shell and T. R. Weikl, *Annual review of biophysics*, 2008, **37**, 289-316.
44. F. Hillenkamp, M. Karas, R. C. Beavis and B. T. Chait, *Analytical Chemistry*, 1991, **63**, 1193A-1203A.
45. M. Karas, D. Bachmann and F. Hillenkamp, *Analytical Chemistry*, 1985, **57**, 2935-2939.
46. J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong and C. M. Whitehouse, *Science (New York, N.Y.)*, 1989, **246**, 64-71.
47. Nobel Media AB, "The Nobel Prize in Chemistry 2002", https://www.nobelprize.org/nobel_prizes/chemistry/laureates/2002/, (accessed 22 May, 2017).
48. C. K. Meng, M. Mann and J. B. Fenn, *Zeitschrift für Physik D Atoms, Molecules and Clusters*, 1988, **10**, 361-368.
49. E. Mirgorodskaya, P. Roepstorff and R. A. Zubarev, *Analytical Chemistry*, 1999, **71**, 4431-4436.
50. L. O. Narhi, C. H. Li, R. Ramachander, J. Svitel and Y. Jiang, in *Molecular Biophysics for the Life Sciences*, eds. N. Allewell, L. O. Narhi and I. Rayment, Springer New York, New York, NY, 2013, DOI: 10.1007/978-1-4614-8548-3_3, pp. 33-90.
51. A. P. Demchenko, in *Ultraviolet Spectroscopy of Proteins*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1986, DOI: 10.1007/978-3-642-70847-3_13, pp. 222-240.

52. C. N. Pace, F. Vajdos, L. Fee, G. Grimsley and T. Gray, *Protein Science*, 1995, **4**, 2411-2423.
53. S. Beychok, *Science (New York, N.Y.)*, 1966, **154**, 1288-1299.
54. C. R. McMillin, W. B. Rippon and A. G. Walton, *Biopolymers*, 1973, **12**, 589-597.
55. D. L. Peterson and W. T. Simpson, *Journal of the American Chemical Society*, 1957, **79**, 2375-2382.
56. D. B. Wetlaufer, *Advances in Protein Chemistry*, 1963, **17**, 303-390.
57. K. Imahori and J. Tanaka, *Journal of Molecular Biology*, 1959, **1**, 359-364.
58. W. Moffitt, *The Journal of Chemical Physics*, 1956, **25**, 467-478.
59. B. A. Wallace, W. R. Veatch and E. R. Blout, *Biochemistry*, 1981, **20**, 5754-5760.
60. G. Weber, *Biochemical Journal*, 1960, **75**, 335-345.
61. F. A. Bovey and S. S. Yanari, *Nature*, 1960, **186**, 1042-1044.
62. M. Ito, *Journal of Molecular Spectroscopy*, 1960, **4**, 106-124.
63. A. Moscowitz, A. Rosenberg and A. E. Hansen, *Journal of the American Chemical Society*, 1965, **87**, 1813-1815.
64. S. P. McGlynn, J. Nag-Chaudhuri and M. Good, *Journal of the American Chemical Society*, 1962, **84**, 9-13.
65. R. Pain, *Current protocols in protein science*, 2005, **Chapter 7**, Unit 7.6.
66. S. Y. Venyaminov and K. S. Vassilenko, *Analytical Biochemistry*, 1994, **222**, 176-184.
67. S. M. Kelly, T. J. Jess and N. C. Price, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 2005, **1751**, 119-139.
68. L. Whitmore and B. A. Wallace, *Biopolymers*, 2008, **89**, 392-400.
69. L. Whitmore and B. A. Wallace, *Nucleic Acids Research*, 2004, **32**, W668-W673.
70. N. Sreerama and R. W. Woody, *Analytical Biochemistry*, 2000, **287**, 252-260.
71. V. Hall, A. Nash and A. Rodger, *Analytical Methods*, 2014, **6**, 6721-6726.
72. V. Hall, M. Sklepari and A. Rodger, *Chirality*, 2014, **26**, 471-482.
73. N. Sreerama, S. Y. U. Venyaminov and R. W. Woody, *Protein Science*, 1999, **8**, 370-380.
74. N. Sreerama and R. W. Woody, *Analytical Biochemistry*, 1993, **209**, 32-44.
75. L. A. Compton and W. C. Johnson, *Analytical Biochemistry*, 1986, **155**, 155-167.
76. P. Manavalan and W. C. Johnson, *Analytical Biochemistry*, 1987, **167**, 76-85.
77. A. Dong, P. Huang and W. S. Caughey, *Biochemistry*, 1990, **29**, 3303-3308.
78. E. Goormaghtigh, V. Cabiaux and J.-M. Ruysschaert, in *Physicochemical Methods in the Study of Biomembranes*, eds. H. J. Hilderson and G. B. Ralston, Springer US, Boston, MA, 1994, DOI: 10.1007/978-1-4615-1863-1_10, pp. 405-450.
79. P. R. Griffiths and J. A. de Haseth, in *Fourier Transform Infrared Spectrometry*, John Wiley & Sons, Inc., 2006, DOI: 10.1002/9780470106310.ch15, pp. 321-348.
80. M. E. Goldberg and A. F. Chaffotte, *Protein Science*, 2005, **14**, 2781-2792.

81. G. Vedantham, H. G. Sparks, S. U. Sane, S. Tzannis and T. M. Przybycien, *Analytical Biochemistry*, 2000, **285**, 33-49.
82. E. Goormaghtigh, J.-M. Ruysschaert and V. Raussens, *Biophysical Journal*, 2006, **90**, 2946-2957.
83. H. Fabian and D. Naumann, *Methods*, 2004, **34**, 28-40.
84. D. Reinstädler, H. Fabian, J. Backmann and D. Naumann, *Biochemistry*, 1996, **35**, 15822-15830.
85. A. Sethuraman and G. Belfort, *Biophysical Journal*, 2005, **88**, 1322-1333.
86. D. M. Byler and H. Susi, *Biopolymers*, 1986, **25**, 469-487.
87. L. G. Hanson, *Concepts in Magnetic Resonance Part A*, 2008, **32A**, 329-340.
88. T. C. Farrar and E. D. Becker, *Pulse and Fourier Transform NMR: Introduction to Theory and Methods*, Elsevier Science, 2012.
89. D. L. Pavia, G. M. Lampman, G. S. Kriz and J. A. Vyvyan, *Introduction to Spectroscopy*, Cengage Learning, 2008.
90. H. J. Dyson and P. E. Wright, *Chemical Reviews*, 2004, **104**, 3607-3622.
91. M. P. Williamson, T. F. Havel and K. Wüthrich, *Journal of Molecular Biology*, 1985, **182**, 295-315.
92. L. Braunschweiler and R. R. Ernst, *Journal of Magnetic Resonance (1969)*, 1983, **53**, 521-528.
93. D. G. Davis and A. Bax, *Journal of the American Chemical Society*, 1985, **107**, 2820-2821.
94. L. Kay, P. Keifer and T. Saarinen, *Journal of the American Chemical Society*, 1992, **114**, 10663-10665.
95. A. Kumar, R. R. Ernst and K. Wüthrich, *Biochemical and Biophysical Research Communications*, 1980, **95**, 1-6.
96. J. Keeler, *Understanding NMR spectroscopy*, John Wiley & Sons, 2011.
97. J. Cavanagh, W. J. Fairbrother, A. G. Palmer Iii, M. Rance and N. J. Skelton, in *Protein NMR Spectroscopy (Second Edition)*, Academic Press, Burlington, 2007, DOI: https://doi.org/10.1016/B978-012164491-8/50006-3, pp. 271-332.
98. V. Tugarinov, R. Sprangers and L. E. Kay, *Journal of the American Chemical Society*, 2004, **126**, 4921-4925.
99. A. Mainz, S. Jehle, B. J. van Rossum, H. Oschkinat and B. Reif, *Journal of the American Chemical Society*, 2009, **131**, 15968-15969.
100. I. Bertini, C. Luchinat, G. Parigi, E. Ravera, B. Reif and P. Turano, *Proceedings of the National Academy of Sciences*, 2011, **108**, 10396-10399.
101. M. Quinternet, J.-P. Starck, M.-A. Delsuc and B. Kieffer, *Journal of Pharmaceutical and Biomedical Analysis*, 2013, **78**, 252-254.
102. A. Jabłoński, *Zeitschrift für Physik*, 1935, **94**, 38-46.
103. J. R. Lakowicz, *Principles of Fluorescence Spectroscopy*, Kluwer Academic/Plenum, 1999.
104. A. P. Demchenko, in *Ultraviolet Spectroscopy of Proteins*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1986, DOI: 10.1007/978-3-642-70847-3_2, pp. 5-26.
105. J. W. Longworth, in *Excited States of Proteins and Nucleic Acids*, eds. R. F. Steiner and I. Weinryb, Springer US, Boston, MA, 1971, DOI: 10.1007/978-1-4684-1878-1_8, pp. 319-484.

106. E. A. Permyakov, *Luminescent Spectroscopy of Proteins*, Taylor & Francis, 1992.
107. R. F. Latypov, D. Liu, K. Gunasekaran, T. S. Harvey, V. I. Razinkov and A. A. Raibekas, *Protein Science*, 2008, **17**, 652-663.
108. M. R. Eftink, in *Topics in Fluorescence Spectroscopy: Volume 6: Protein Fluorescence*, ed. J. R. Lakowicz, Springer US, Boston, MA, 2000, DOI: 10.1007/0-306-47102-7_1, pp. 1-15.
109. M. Kerker, *The scattering of light and other electromagnetic radiation*, Elsevier, 2016.
110. H. C. Hulst and H. C. van de Hulst, *Light Scattering by Small Particles*, Dover Publications, 1957.
111. N. A. Clark, J. H. Lunacek and G. B. Benedek, *American Journal of Physics*, 1970, **38**, 575-585.
112. D. E. Koppel, *The Journal of Chemical Physics*, 1972, **57**, 4814-4820.
113. U. Nobbmann, FAQ: Peak size or z-average size – which one to pick in DLS?, http://www.materials-talks.com/blog/2014/07/10/faq-peak-size-or-z-average-size-which-one-to-pick-in-dls/, (accessed 31 May, 2017).
114. Malvern Instruments, Dynamic Light Scattering: An Introduction in 30 Minutes, https://www2.warwick.ac.uk/fac/cross_fac/sciencecity/programmes/internal/themes/am2/booking/particlesize/intro_to_dls.pdf, (accessed 22 Jun, 2017).
115. R. Furth and A. Cowper, *Investigations on the Theory of the Brownian Movement by Albert Einstein*, Dover, New York, 1956.
116. A. A. Wakankar and R. T. Borchardt, *Journal of Pharmaceutical Sciences*, 2006, **95**, 2321-2336.
117. J. Patel, R. Kothari, R. Tunga, N. M. Ritter and B. Tunga, *BioProcess International*, 2011, **9**, 20-31.
118. T. J. Kamerzell, R. Esfandiary, S. B. Joshi, C. R. Middaugh and D. B. Volkin, *Advanced Drug Delivery Reviews*, 2011, **63**, 1118-1159.
119. R. Nayar and M. Mosharraf, in *Formulation and Process Development Strategies for Manufacturing Biopharmaceuticals*, John Wiley & Sons, Inc., 2010, DOI: 10.1002/9780470595886.ch13, pp. 307-328.

Abstract

Although protein structure has been studied for many decades, it remains the case that we cannot state with confidence whether two samples have the same molecular structure, particularly in solution. The increasing number of biosimilar biopharmaceutical drugs that are being tested means this is not an academic exercise. In this work we consider how five well-established techniques: dynamic light scattering (DLS), circular dichroism (CD), Fourier transform infrared spectroscopy (FT-IR), nuclear magnetic resonance spectroscopy (NMR), and molecular modelling can be combined to provide information about the supposedly well-understood protein insulin. A goal of this work was to establish a systematic means of detecting differences between insulin samples as a function of pH, temperature, and the presence or absence of zinc, all of which are known to change the oligomerisation state and to affect molecular structure. We used the recently developed Secondary Structure Neural Network (SSNN) circular dichroism algorithm to facilitate analysis of the CD spectra as well as a retrained version for analysis of the FT-IR spectra.

Chapter 2
Biophysical Characterisation of Insulin for Structure Comparison

2.1. INTRODUCTION

Since 1921 when insulin was first extracted from dogs and purified, the advances in insulin's history were slow in pace but tremendous in importance leading to the award of several Nobel Prizes to the respective discoverers.[1,2] In 1955 insulin was the first protein to be fully sequenced, and subsequently it was the first human protein to be chemically synthesised in 1963. With the developments in X-Ray crystallography, Dorothy Hodgkin and her team solved the three-dimensional structure of insulin in 1969.[3] In 1978, insulin was the first human protein to be genetically engineered by Genentech, and launch of recombinant human insulin as the first biopharmaceutical drug (Eli Lilly & Co.'s Humulin) followed in 1982.[4]

Despite the fact that insulin has been extensively studied since the 1920s, a lot has to be explored and it is still under research. Insulin products have recently lost their patent protection and several biosimilar products as well as new drugs have been developed or are under development for diabetes treatment.[5] The main focus is on ultra-rapid acting insulins, new formulations with prolonged duration of action (such as insulin degludec),[6] innovative administration methods (oral or inhaled) and new devices for insulin delivery (infusion pumps).[7] The case where Marvel Ltd withdrew its application for a marketing authorisation for a biosimilar insulin renders even more significant the need for high standards of quality control.[8]

In this chapter we focus on bovine insulin as a case study, considering what information can be extracted from five of the well-established techniques (CD, NMR, DLS, FT-IR and molecular modelling) often used to characterise protein structure. We have complemented room temperature data collection with temperature dependence which enables us to probe subtle differences in structure that may not be apparent in room temperature data. We chose insulin as our model system for this study since it plays a key role in the regulation of blood glucose levels, exists in different forms, and many of the studies in the past years have used circular dichroism (CD),[9-11] nuclear magnetic resonance (NMR)[12-18] and Fourier transform infrared spectroscopies (FT-IR),[19-24] dynamic light scattering (DLS),[25-29] X-Ray crystallography,[30-33] mass

spectrometry (MS)[34,35] and other techniques. Insulin is often described as a peptide hormone since it consists of two polypeptide chains A and B, with 21 and 30 amino acids respectively, held together by two disulfide bridges Cys (A7)–Cys (B7) and Cys (A20)–Cys (B19), while a third intrachain disulfide bond links cysteines A6 and A11.[36-42] Usually, chain A folds into two α-helical segments A2–A8 and A13–A19 connected by an extended section, while chain B folds into two extended segments connected by a central α-helical segment at B9–B22[31,43] (Figure 2.1). Bovine insulin differs in sequence from human insulin at A8 (human: Thr; bovine: Ala), A10 (human: Ile; bovine: Val) and B30 (human: Thr; bovine: Ala), resulting in a slightly more compact structure for the first.[30]

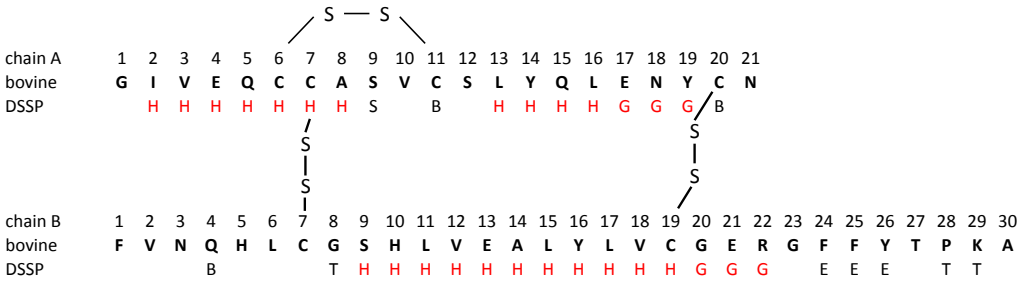


Figure 2.1: Sequence and secondary structure of bovine insulin. The disulfide bridges are shown and the amino acids that belong to helices are annotated in red. H = α-helix, G = 3/10-helix, S = bend, B = β-bridge, T = turn, E = β-strand. The secondary structure is according to the DSSP annotation for the entry with PDB ID 4E7T.

One of the main characteristics of insulin is its ability, under different conditions, to exist as:

- monomers (pH 10–11) or
- dimers (acidic pH, with the C-terminal portions of two adjacent B chains forming anti-parallel strands)[44] or
- tetramers and hexamers (neutral pH, with the hexamer being three dimers assembled around 2 Zn^{2+} ions to bury hydrophobic surfaces).[16,27,45-48]

In hexamer formation (Figure 2.2), three identical dimers assemble around 2 Zn^{2+} ions, with each metal ion coordinated in an octahedral arrangement to three His (B10) side chains and three water molecules.[49,50] In this way, more hydrophobic surface areas are buried in the hexamer's interior and therefore, in the organism, the hexamer serves for storage of insulin molecules in membrane enclosed vesicles rich in zinc, where they are protected from physical and chemical degradation.[51,52] During transfer from vesicles into the blood stream, the Zn^{2+}-insulin hexamers rapidly break down to

monomers (the active form),[53,54] favoured by the significant dilution that takes place (insulin concentration in vesicles is ~21 mM; in circulation 57–280 pM), by the change in pH from 5.5 to 7.4, which results in strong repulsion of the six Glu (B13) carboxylates, and also by the limited availability of zinc ions which show higher affinity to albumin in serum.[52,55] Monomeric units within the different oligomerisation states have similar but not identical structures, with the hexamer being the most stable and slightly more helical than any other oligomerisation state.[56] In the present study, EDTA was used to remove Zn-binding in the case of neutral and basic pH,[55] while at acidic pH, the His (B10) side chains are protonated and therefore unable to coordinate metal ions.[57]

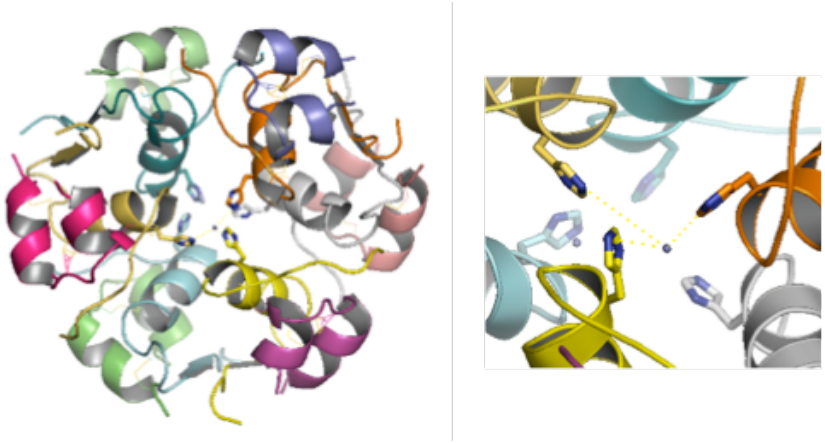


Figure 2.2: *Left*: Three dimers assemble around 2 Zn^{2+} ions in hexamer formation. Each chain of the insulin hexamer is represented with a different colour. *Right*: Coordination between a Zn^{2+} ion and three His (B10) side chains. For the representation the structure with PDB ID code 4E7T was used and PyMOL.[58]

Insulin is also one of the proteins that are known to be involved in amyloidosis. Such proteins lead to the formation of intermolecular β-sheets that result from aggregation of the unfolded protein under stressful conditions (*e.g.* increased temperature). These structures gradually evolve resulting in the formation of fibrils (amyloid-like fibrils), which are responsible for many neurodegenerative diseases such as Parkinson's and Alzheimer's,[19] while insulin aggregates have been found responsible for cutaneous allergy.[59] Aggregation and fibrillation studies are therefore important in order to understand the mechanism and conditions under which it occurs. Bovine insulin, which was used in the present study, is more prone to form fibrils than human insulin.[60]

The variation of structure with conditions (monomer, dimer, hexamer depending on pH and presence or absence of zinc) is important in the formulation of

commercial insulin products, as insulin's oligomerisation state is directly connected with its biological activity (see commercial formulations).[51-55] In the present study, our focus is on the relationship between oligomerisation state, stability, secondary structure and aggregation or fibrillation (when apparent from the data) and the simplest analytical methodologies that can be used to probe this. The behaviour of the monomer, dimer and hexamer as a function of temperature was measured by four well-established experimental techniques—DLS, CD, NMR and FT-IR spectroscopy—and was simulated through molecular dynamics (MD) simulations to aid interpretation of the experimental data.

Commercial formulations

In the same way that a change in the substitution of a small molecule can affect its specific action (*e.g.* vinblastine/vincristine), a mutation in the amino acid sequence of a biological molecule can affect its structure and in consequence its action. Different products (analogues) are often found under the same name of the native protein, and a well-known example is that of insulin. Changes such as replacement, addition of specific amino acids or attachment of a fatty acid chain on the insulin molecule (Figure 2.3) have led to numerous products that differ in their pharmacokinetics. The resulting analogues therefore are separated into four types (rapid-, short-, intermediate- and long-acting) depending on their onset, peak and duration of action (Table 2.1).

chain A	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
human	**G**	**I**	**V**	**E**	**Q**	**C**	**C**	**T**	**S**	**I**	**C**	**S**	**L**	**Y**	**Q**	**L**	**E**	**N**	**Y**	**C**	**N**
glargine																					G

chain B	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
human	**F**	**V**	**N**	**Q**	**H**	**L**	**C**	**G**	**S**	**H**	**L**	**V**	**E**	**A**	**L**	**Y**	**L**	**V**	**C**	**G**	**E**	**R**	**G**	**F**	**F**	**Y**	**T**	**P**	**K**	**T**		
glulisine			K																										E			
lispro																												K	P			
aspart																												D				
glargine																														R	R	
detemir																													C14 fatty acid chain			
degludec																													side-chain of E + C16 fatty acid			

Figure 2.3: Sequences of native human insulin and its analogues. For the analogues only the sites of the altered amino acids are shown, while the rest of the sequence (blank) remains the same as in the native protein.

A key property in order to achieve rapid or prolonged action is the oligomerisation state of insulin. As can be easily observed from Table 2.1, the majority of the formulations contain a source of Zn ions ($ZnCl_2$ or ZnO) and are kept in neutral pH, conditions that would favour the formation of hexamers in the case of native human insulin at the used concentration (~3.5 mg/mL). However, insulin analogues in fast-acting formulations (such as insulin glulisine, lispro and aspart) have a reduced tendency to form hexamers[61] and therefore are more rapidly absorbed compared to

regular insulin. On the opposite end, the addition of a fatty acid at the C-terminal of chain B (insulin detemir, degludec) increases self-association of the drug molecules resulting in a slow systematic absorption. Similarly, insulin glargine leads to the formation of microprecipitates when neutralised after injection, leading to a slow release of insulin molecules.[62] Finally, the duration of action of formulations containing recombinant human insulin (short) and protamine (intermediate action) largely depends on several factors (*e.g.* dose, site of injection, physical activity) and varies in different individuals or at different times in the same individual. Thus, combination of a short- and intermediate- acting insulin in fixed ratios is quite common (pre-mixed formulations) offering more effective coverage in basal (fasting control) and bolus (mealtime carbohydrate control) insulin needs.[63]

Table 2.1: Commercial formulations of insulin. The mentioned quantities for the active substance and the inactive ingredients concern preparations with strength of 100 units/mL. According to WHO 1 IU (International Unit) of human insulin is the activity contained in 0.03846 mg of the international standard for human insulin.[64] NPH = Neutral Protamine Hagedorn.

Type (acting)	Brand	Recombinant human analogue	Appearance	mg/mL	Inactive ingredients per mL	pH
Rapid	Apidra[1]	insulin glulisine	clear, colourless solution	3.49	3.15 mg m-cresol 6 mg tromethamine 5 mg NaCl 0.01 mg Tween 20 water for injection HCl, NaOH for pH adjustment	7.3
	Humalog[2]	insulin lispro	clear, colourless solution	3.47	0.0197 mg Zn 3.15 mg m-cresol 16 mg glycerin 1.88 mg Sodium Phosphate dibasic	7.0-7.8
	NovoRapid[3]	insulin aspart	clear, colourless solution	3.50	Zinc chloride glycerol phenol m-cresol NaCl Disodium phosphate dihydrate water for injection HCl, NaOH for pH adjustment	
Short (Regular)	Humulin R[4] (Novolin R) (Actrapid) (Insuman)	identical to human insulin	clear, colourless solution	-	0.015 mg Zn 2.5 mg m-cresol 16 mg glycerin water for injection HCl, NaOH for pH adjustment	7.0-7.8
Intermediate (NPH)	Humulin N[5] (Novolin N) (Insulatard) (Protaphane)	regular insulin + protamine (human isophane suspension)	cloudy suspension	-	0.025 mg Zn 0.35 mg protamine sulfate 16 mg glycerin 3.78 mg dibasic sodium Phosphate 1.6 mg m-cresol 0.65 mg phenol water for injection HCl, NaOH for pH adjustment	7.0-7.5
Long	Lantus[6] (Abasalgar)	insulin glargine insulin glargine (Biosimilar)	clear aqueous solution	3.64	0.03 mg Zn 2.7 mg m-cresol 20 mg glycerol 85% 0.020 mg Tween 20 water for injection HCl, NaOH for pH adjustment	4.0
	Levemir[7]	insulin detemir	clear, colourless solution	14.20	0.0654 mg Zn 2.06 mg m-cresol 16 mg glycerin 1.80 mg phenol 0.89 mg disodium phosphate dihydrate 1.17 mg NaCl HCl, NaOH for pH adjustment	7.4
	Tresiba[8]	insulin degludec	clear, colourless solution	-	0.0327 mg Zn 19.6 mg glycerol 1.50 mg phenol 1.72 mg m-cresol water for injection HCl, NaOH for pH adjustment	7.6

Information for the types of insulin were collected from[65-68]
Information about the ingredients were collected from the products' labels. The following links were accessed on 17 Jan, 2017:

[1] https://www.drugs.com/pro/apidra.html
[2] http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/020563s115lbl.pdf
[3] http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Product_Information/human/000258/WC500030372.pdf
[4] http://www.accessdata.fda.gov/drugsatfda_docs/label/2011/018780s120lbl.pdf
[5] http://pi.lilly.com/us/HUMULIN-N-USPI.pdf
[6] http://www.accessdata.fda.gov/drugsatfda_docs/label/2009/021081s034lbl.pdf
[7] http://www.accessdata.fda.gov/drugsatfda_docs/label/2012/021536s037lbl.pdf
[8] http://www.accessdata.fda.gov/drugsatfda_docs/label/2015/203314lbl.pdf

2.2. MATERIALS AND METHODS

2.2.1. Materials

Insulin from bovine pancreas (I6634) and ethylenediamine tetraacetic acid disodium salt dihydrate (E5134) (Na_2EDTA) were purchased from Sigma-Aldrich. Hydrochloric acid (36%; BDH) was purchased from VWR International, sodium hydroxide pellets and sodium dihydrogen orthophosphate dihydrate were purchased from Fisher Scientific, and disodium hydrogen orthophosphate (71643) was from Fluka Biochemika. Latex spheres (3100A) were from Thermo Fisher scientific. Deuterium oxide was from Sigma-Aldrich (151882), and ultrapure water was used (18 MΩ·cm at 25 °C) in all cases. The 0.1 M HNO_3 for ICP-OES sample preparation was purified in-house by sub-boiling point distillation, and for calibration, zinc ICP/DCP 356743 and sulfur ICP/DCP 356603 standards (Fluka) were used.

2.2.2. Spectroscopic data collection

A Jasco V-660 ***UV-Vis*** or a NanoPhotometer P300 UV-Vis Spectrophotometer was used for measuring insulin concentrations using a molar absorptivity value of ε_{278} = 6080 $M^{-1}cm^{-1}$.[69] For the UV-Vis calibration curve, the concentration of the prepared by weight sample was determined by its absorbance as 1.75 mg/mL, which was used as the starting point. Insulin was subjected to consecutive dilutions and each spectrum was acquired following its respective blank with the instrument set to automatic baseline correction. All data were recorded in a 1 cm quartz cell, from 250–310 nm, with a data interval of 0.2 nm, bandwidth 2.0 nm and scanning speed 100 nm/min. In Figure 2.4, the marker points represent the expected concentrations (following dilutions of the 1.75 mg/mL sample) and a straight line was fit to the data.

DLS data were collected on a Malvern Zetasizer Nano-series with a laser wavelength of 633 nm and a detection angle of 173˚. A quartz cuvette (ZEN2112) QS 3.00 mm from Malvern Instruments was used for the measurements. Averages over 6 measurements at each temperature for each sample were taken and are presented in terms of intensity and number values. Number values present the number of particles of each size, thus emphasizing smaller particles, whereas intensity values scale by the volume, thus emphasizing the presence of larger particles.

CD spectra were collected on a JASCO J-1500 CD spectropolarimeter using a PTC-510 Peltier thermostatted cell holder. The CD spectra (mdeg) were baseline-corrected by subtracting the corresponding buffer blank from each spectrum at all temperatures. The baseline-corrected spectra were zeroed by subtracting the averaged

value from 260–250 nm (mdeg) from the whole spectrum. The data were converted into Δε per amino acid residue units, the wavelength step was changed to 1 nm and data from 240 nm to 190 nm were used as the input format required for the secondary structure analysis by SSNN[70] and SELCON.[71] The CD spectra with high concentration (~1 mM) were normalised according to the negative peak at ~208.0 nm of the low concentration (~17 μM) spectra in order to facilitate qualitative comparison of the CD spectra at the two different concentrations.

IR spectra were collected on a single reflection ZnSe Attenuated Total Reflectance (ATR) accessory, PIKE MIRacle with a high-pressure clamp, on a JASCO FT/IR-4200 spectrometer. The sample was heated in the water bath of a Hei-VAP rotary evaporator and the temperature was checked with a Digi-Thermo thermometer (−55°C to +148°C). Air was recorded as a background. The IR spectra (Absorbance units) of the dried sample layers were manually corrected by subtracting a water vapour spectrum. The corrected spectra were zeroed in the region of 1800 cm^{-1} and normalised by dividing by the maximum absorbance value between 1600 to 1700 cm^{-1}. The SSNN software was retrained with a reference FT-IR data set of 47 proteins with known secondary structures from crystallography data, provided by Rina Dukor (private communication), and the Amide I band was tested for its secondary structure content. The second derivative was calculated for the area 1400–1800 cm^{-1} in order to show the exact positions of the peaks, with a five data point window and without smoothing.

Protein (*via* Sulfur) and zinc concentrations were determined using an Inductively Coupled Plasma Optical Emission Spectrometer (***ICP-OES***; PerkinElmer Optima 5300 DV, model S10) by measuring S at 180.669 and 181.975 nm and Zn at 206.200 and 213.857 nm. The measurements were conducted by Dr. Hasan Tanvir Imam. Plasma operating conditions were: Argon (Ar) flow rate 13.0 L/min, auxiliary gas flow rate 0.2 L/min, nebulizer flow rate 0.8 L/min and RF power 1300 W. For calibration, 6 Zn and S standards of 0.2, 0.5, 0.7, 1.0, 2.0 and 5.0 ppm were prepared gravimetrically in ultra-pure 0.1 M HNO_3. Samples were prepared using the same 0.1 M HNO_3, which was also used as a blank. Typically, correlation coefficients of better than 0.9998 were obtained.

For all ***NMR experiments***, the water peak was suppressed using double pulsed field gradient spin echo for excitation sculpting (DPFGSE).[72] All spectra were Fourier-transformed, phased and baseline-corrected in Topspin 1.3. ***1D ^{1}H-NMR experiments*** were carried out on a Bruker AV III-600 spectrometer operating at 600.13 MHz for ^{1}H.

Exponential multiplication was used for apodisation and line broadening 0.5 Hz; no zero-filling was applied. The chemical shift of the residual HDO was corrected for its temperature-dependence and used as internal reference.[73] The line widths of selected peaks, which could be easily followed across the increasing temperature, were measured in Topspin 1.3 at half height of the peaks. ***2D-NMR experiments*** were recorded on a Bruker AV II-700 spectrometer operating at a ^{1}H frequency of 700.24 MHz and a ^{13}C frequency of 176.08 MHz. Data matrices of 4096 (t2 dimension) by 440 (t1 dimension) with acquisition times of 0.18 s and 0.02 s (in the respective dimensions) were collected as a sum of 32 scans for the ^{1}H–^{1}H chemical shift-correlated spectra (NOESY and TOCSY), and the spectra were processed with 2048 points in each dimension of the frequency domain. For the TOCSY spectrum, homonuclear Hartman-Hahn transfer using the MLEV17 sequence[74] with a mixing time of 65 ms was used. For the NOESY experiment, a mixing time of 125 ms was used. The ^{1}H–^{13}C correlated spectrum (HSQC) was a sum of 192 scans over data matrices of 2048 (t2) by 128 (t1), using 2D H-1/X correlation via double inept transfer and Echo/Antiecho-TPPI (time-proportional phase incrementation) acquisition with heteronuclear decoupling during acquisition.[75-77] The spectral widths were 16 ppm and 166 ppm and relaxation delay was 1 s. The spectra were processed with 2048 (f2 dimension) and 1024 (f1 dimension) points. The squared shifted sinebell function was used for apodisation of all 2D spectra.

The ***balance*** used was a Mettler Toledo XP2U and the ***pH meter*** Mettler Toledo Seven Compact pH/Ion S220 InLab Nano Sensors. ***Refractive indices*** were measured on a Bellingham+Stanley Abbe60/DR Refractometer. The ***centrifuges*** were a Fisher Scientific microcentrifuge 7200g and a SIGMA 1–14K. A Merck Millipore Direct-Q system was used for ***water purification***.

2.2.3. Temperature dependent measurements

For all ***CD temperature-dependent experiments***, the same parameters were used: wavelength range 260–180 nm, 1 s data integration time (DIT), 2 nm band width, data pitch 0.2 nm, 100 nm/min continuous scanning speed. Data were obtained every 5 °C from 20 °C to 110 °C, at a ramp rate of 2.5 °C/min with 300 s equilibration time once the desired temperature was reached. CD spectra were also recorded at the conclusion of the temperature experiment, 5–8 hours after the samples had been cooled down to 20 °C. Each spectrum is the average of 12 accumulations. A 0.1 cm path length cuvette was used for 0.1 mg/mL insulin solutions. For the high concentration solutions (~1 mM) a 0.01 cm fixed path length round cuvette and a demountable cuvette with a

nominal path length of 0.001 cm were used. The path length of the demountable cuvette was measured by using the absorbance of 0.2 M potassium chromate and the Beer-Lambert Law with an extinction co-efficient of 4830 $M^{-1}cm^{-1}$ at 372 nm. The length was determined to be 0.00208 ± 0.00008 cm, and the value 0.00208 was used in the analysis.

For the ***1D ^{1}H-NMR temperature-dependent experiments*** the acquisition time was either 0.7 or 1.2 seconds, a relaxation delay of 1 s, and a 90˚ pulse width of 8.05 μs were used. The final spectra were the sum of 128 scans. An equilibration time of 5 minutes was allowed after the desired temperature was reached. The chemical shift of the water resonance was determined at each temperature and the excitation frequency (centred on the water resonance) was adjusted accordingly. Spectra were acquired from 20–75 °C, with a 5 °C temperature interval.

For the ***DLS experiments***, the viscosity was estimated as described by Gilroy *et al.*[78] For the *neutral and basic samples* (Table 2.2), measurements of solutions consisting of 1 μL of polystyrene latex spheres of known size (100 nm) in 99 μL of blank solution at each temperature were run, while for the *acidic sample* the viscosity of water was used, as latex spheres appeared to aggregate in these conditions due to increased hydrophobicity.[79] In a temperature range from 20–75 °C with a 5 °C temperature interval, 6 measurements were recorded for each temperature. 300 s were allowed as the equilibration time after the desired temperature was reached. Each measurement consisted of 6 runs, with 60 s total duration. In the case of the *hexamer control sample*, the viscosity was estimated by measuring a solution consisting of 1 μL of latex spheres in 25 μL of blank at each temperature. Each measurement consisted of 10 runs, with 200 s total duration time, and 180 s were allowed as the equilibration time.

For the ***FT-IR temperature experiment,*** the instrument was purged with N_2 for at least 1 hour before recording the first spectrum and continuously throughout the measurements. The sample was heated for 20 minutes at each temperature, in a range from 20–95 °C with a 5 °C interval. Spectra were recorded at 25 °C, after being heated for 20 minutes at 75 °C and for one hour at 95 °C. For the data collection, 5 μL of sample were applied on the crystal and dried with CO_2-free air. The procedure was repeated 4 times for each sample resulting in a 4-layer solid. A spectrum of solid fibres was also recorded for a sample that was prepared and treated in the same way and was left at room temperature for 5 months. All data were collected at room temperature,

using the TGS (triglycine sulfate) detector and in a wavelength range from 400–4000 cm^{-1}. For each spectrum 1000 scans with 4 cm^{-1} resolution were co-added.

Table 2.2: Corrected values for viscosity at each temperature and refractive indices for each one of the blank solutions, obtained as described in the text. The values presented for the blank at pH 2.5 are the same as water values.

Temperature (°C)	Blank pH 2.5	Blank pH 7.7	Blank pH 7.6, EDTA	Blank pH 10.9	Blank pH 10.5, EDTA	Blank pH 8.0, 100 mM NaCl
20	1.003	1.018	0.991	1.075	1.038	0.988
25	0.887	0.936	0.906	0.945	0.939	0.874
30	0.792	0.852	0.776	0.790	0.831	0.785
35	0.714	0.759	0.683	0.724	0.710	0.701
40	0.649	0.675	0.648	0.673	0.683	0.637
45	0.595	0.616	0.592	0.621	0.631	0.593
50	0.548	0.580	0.533	0.566	0.556	0.552
55	0.508	0.522	0.491	0.520	0.527	0.514
60	0.471	0.533	0.455	0.476	0.479	0.479
65	0.438	0.477	0.433	0.478	0.461	0.440
70	0.407	0.453	0.398	0.444	0.410	0.414
75	0.378	0.456	0.375	0.425	0.400	0.385
Refractive Index	1.3330	1.3334	1.3332	1.3345	1.3335	1.3340

2.2.4. Sample preparation

For the ***UV-Vis calibration***, 2.151 mg of insulin were weighed and dissolved in 1.076 mL of 0.3 M NH_4OH. The pH was measured as 10.5.

For the ***CD experiments***, the different samples were prepared as follows:

⇒ *Low pH sample*: insulin was initially dissolved in HCl (0.1 M). Sodium phosphate buffer (100 mM, pH 7.0) and water were added. The pH was further adjusted with HCl (0.1 M) and the final volume was adjusted with the addition of water, resulting in an insulin solution of ~0.1 mg/mL in 10 mM sodium phosphate buffer and pH 2.8.

⇒ *Neutral pH samples*: insulin was dissolved in 0.1 M NaOH; sodium phosphate buffer was added and the solution was split into two parts. EDTA (1 mM) was added to one of the aliquots (to complex zinc), aiming at a concentration 1.5 times that of the concentration of insulin in the sample. Water was added resulting in insulin solutions of ~0.1 mg/mL in 10 mM sodium phosphate buffer. The pH of both samples was 7.3.

⇒ *Basic pH samples*: two samples were prepared as for the neutral samples except that pH was adjusted with 0.1 M NaOH to pH 10.6 in 10 mM sodium phosphate buffer.

For the *high concentration solutions* the preparation was the same as for the respective EDTA-free samples, resulting in insulin solutions of 1.18 mM at pH 2.3, 0.90 mM at pH 7.5 and 0.92 mM at pH 10.9.

All samples were left to equilibrate for 30 mins and the pH was checked again before starting the experiment. For all samples, identical control solutions omitting the protein (blank) were prepared for baseline correction of CD spectra.

Samples* for *DLS were prepared in the same way as samples for CD spectroscopy, but with a 50-fold increase in all concentrations. The resulting pH values were: 2.5, 7.7 and 10.9 for the EDTA-free basic sample and 10.5 for the EDTA-containing basic sample. The *hexamer control sample*[25] was prepared in the same way as the neutral EDTA-free sample, with the only difference being the addition of NaCl (1 M), resulting in an insulin solution of ~8 mg/mL in 10 mM sodium phosphate buffer, 100 mM NaCl and pH 8.0. The refractive indices of the blank samples were measured. All samples and blanks were centrifuged with a centrifugal force of 6,000 rpm for 10 minutes and filtered through 0.2 μm syringe filters (Thermo Scientific) directly into a Quartz cuvette prior to the measurement.

For the ***^{1}H-NMR experiments***, the samples were prepared as for the DLS experiments, but solutions contained 10% D_2O for the field-frequency lock.[48] The resulting pH values (determined with the pH meter calibrated using buffers in pure H_2O solutions) were: 2.5, 7.4 and 10.8 for the EDTA-free basic sample and 10.7 for the EDTA-containing basic sample. The ***2D-NMR experiments*** (NOESY, TOCSY, ^{13}C-HSQC) were recorded for an acidic sample (pH 2.5) prepared in the same way, and at a temperature of 25 °C.

For the ***FT-IR temperature experiment*** the sample was prepared in the same way as the *hexamer control sample* for the DLS experiment, differing slightly in insulin and buffer concentration (10 mg/mL and 50 mM respectively) and a pH of 8.2. Solid fibres were obtained from a similar sample that was left at room temperature for 5 months.

For the ***ICP-OES sample preparation***, insulin was initially dissolved in 1 M NH_4OH. 20 mM ammonium bicarbonate buffer (pH 8.0) and water were added, resulting in a final insulin solution of 1 mg/mL (theoretical concentration) in 10 mM ammonium bicarbonate and 0.1 M NH_4OH. The sample was separated into 100, 200, 300 and 500 μL aliquots, and each was diluted to a final volume of 3 mL using ultra-pure 0.1 M HNO_3.

2.2.5. Simulated Annealing (SA)

Simulated annealing molecular dynamics (SA-MD) simulations were performed using the AMBER 12.0 package. Each system was prepared by using the Generalised Born implicit solvent model.[80] The Particle Mesh Ewald (PME) method was employed in all of the simulations.[81] During each simulation, all bonds to H atoms were fixed and all other bonds were constrained to their equilibrium values by applying the SHAKE algorithm.[82] A cut-off radius of 9 Å was used. The minimisations were performed for 10000 cycles. The temperature was controlled by applying the Langevin algorithm, while the pressure was controlled by the isotropic position scaling protocol used in AMBER.[83] Three separate simulations of insulin monomer, dimer and hexamer were performed. The initial PDB (Protein Data Bank) structures for monomer, dimer and hexamer were extracted from the PDB ID code 4E7T. The pH was set to 10.8, 2.5 and 7.5 for insulin monomer, dimer and hexamer systems respectively with appropriate states of protonation, deprotonation and neutralisation of the amino acid residues present in each insulin system, as established using Accelrys discovery studio 2.5. The temperature range used was 293–383 K. At the first step after minimisation, the systems were heated to 293 K and allowed to relax slightly. A total run of 1.68 ns of MD simulated annealing with the time and temperature profile given in Table 2.3 was performed for all insulin systems.

For the ***secondary structure determination***, the DSSP (Dictionary of protein Secondary Structure)[84] module of AMBER was used and three snapshots of the annealing trajectories at 20, 50, 65, 75 and 110 °C were analysed for each system. The results presented are the average over three snapshots at 1/3, 2/3 and last trajectory of each temperature. α-helix and 3/10-helix were combined to give the total α-helix content; β-strand was used for the total β-sheet content; turns were given by addition of turns and bends; and "other" is the sum of β-bridge, π-helix and other, according to Micsonai *et al*. and Sreerama *et al.*[85,86] The images from the simulations were generated using Swiss-PdbViewer.[87]

Table 2.3: Time and temperature profile of the SA-MD simulations for all insulin systems.

Steps	Temperature (K)	Time (fs)
Annealing Step 1	0–293	500
Annealing Step 2	293	60000
Annealing Step 3	293–323	60000
Annealing Step 4	323	240000
Annealing Step 5	323–338	30000
Annealing Step 6	338	120000
Annealing Step 7	338–348	20000
Annealing Step 8	348	80000
Annealing Step 9	348–383	70000
Annealing Step 10	383	800000
Annealing Step 11	383–293	100000
Annealing Step 12	293	100000
Total Time		1680500 fs = 168 ps = 1.68 ns

2.3. Concentration Determination

2.3.1. ICP-OES measurements

According to the manufacturer's specification sheet, the Zn content in the product is ≤1.0 % (on a dry basis). ICP-OES was conducted in order to verify the protein concentration and the amount of Zn present. The measured sulfur content was divided by the number of cysteines in the monomer resulting in 0.80 mg/mL insulin concentration, [P]. The ratio of the measured zinc content to the protein's concentration was found to be [Zn]/[P]=0.65 (0.7 % on a dry basis). The sample was also checked by UV-Visible absorbance spectroscopy (UV-Vis); this gave the same result. The significant and consistent difference observed between the weighed theoretical concentration (1 mg/mL) and the experimentally measured concentration (0.80 mg/mL) of the tested sample is probably due to the amount of water that solid insulin might contain. After confirming that the two methods were in agreement, we decided to use UV-Vis as the main method for estimating the protein's concentration and to make that easier, the following calibration curve was measured.

2.3.2. UV-Vis spectroscopy

As mentioned above, the experimentally measured concentration (1.75 mg/mL) is persistently lower than the theoretical concentration of the originally weighed insulin (2 mg/mL) by a factor of ~0.85 on average. The calibration curve (Figure 2.4) provided data on the empirical relationship between insulin's concentration and absorbance, and allowed for a quick estimation of the protein's concentrations in solution. Although the used solvent was different than the conditions used in the rest of the experiments (see sample preparation)—which can potentially affect the instrumental response and subsequently the derived equation for the protein's concentration—the curve can still give a good approximation and the equation is very close to the Beer-Lambert Law when ε_{278} = 6080 $M^{-1}cm^{-1}$, for 1 cm path length and after conversion from molar to mass concentration for insulin (5733.49 g mol^{-1}) is applied:

$$A = \varepsilon c L \Rightarrow A = \varepsilon \frac{\rho}{\mathrm{M}} L \Rightarrow A = 1.0604\rho$$

where A: absorbance, ε: extinction coefficient, c: concentration (M), L: path length (cm), ρ: concentration (g L^{-1}) and M: molecular weight (g mol^{-1}).

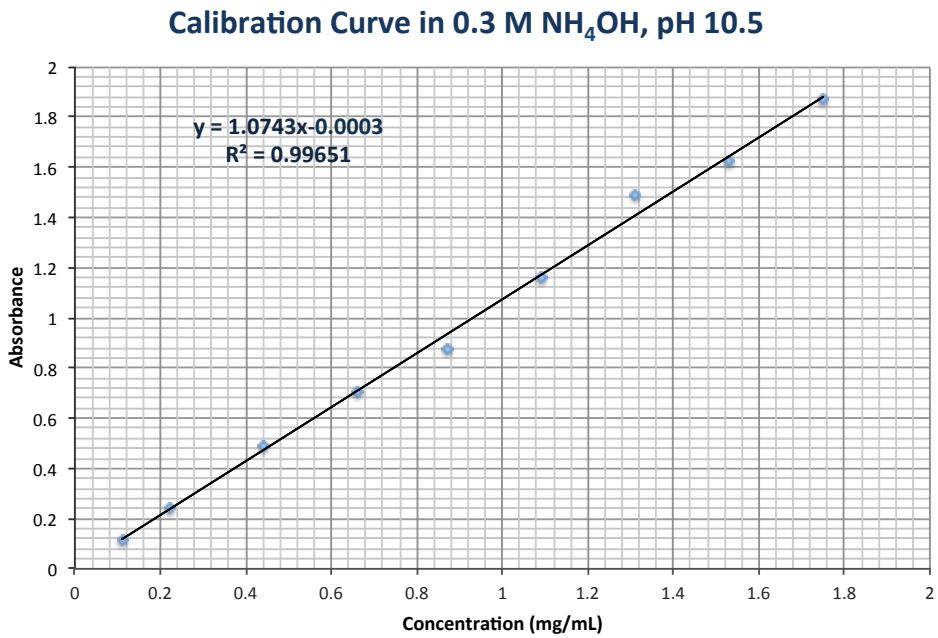


Figure 2.4: The concentration of insulin in 0.3 M NH_4OH at pH 10.5 was calibrated in relation to the UV-Vis absorbance at 278 nm in a 1 cm cuvette. The equation and the R-squared value are shown in the chart.

2.4. DLS SPECTROSCOPY

Insulin assembly and aggregation is pH and concentration dependent.[27,88] Based on the X-ray structure,[36] the monomer, dimer, tetramer and hexamer diameters with one layer of water have been calculated previously (in the literature) as 3.1, 3.7, 4.5, and 5.7 nm, respectively.[25,88,89] In our dynamic light scattering (DLS) experiments at 1 mM protein, three main populations (Figure 2.5i,ii, Table 2.4) were observed below 40 °C. At 20 °C (Figure 2.5i), an equilibrium of monomers and dimers (intensity values 3.2–3.3 nm) is dominant at acidic and basic pH, whereas tetramers (4.8 nm) are predominant at neutral pH in the presence of EDTA (no Zn), and hexamers (5.3 nm) are favoured at neutral pH in the presence of Zn, possibly in equilibrium with smaller oligomers as judged by the relatively small size value. The acidic and neutral samples presented mid-range polydispersity (polydispersity index >0.1), while the basic samples appeared to be highly polydisperse (polydispersity index >0.4). All solutions also contained very high molecular weight species (data not shown). The hexamer control sample was prepared in conditions that strongly favoured the hexamer formation (1.4 mM insulin, 100 mM NaCl, pH 8.0).[25] In the temperature range from 35–65 °C it appeared to be nearly monodisperse, showing one main population to be present below 70 °C with hydrodynamic diameter (intensity values) of 5.5–6.1 nm. Larger aggregates appeared at 75 °C, while the sample had probably evaporated as the intensity decreased significantly (Figure 2.5iii, bottom right). The size distributions at 20 °C and from 60–75 °C are shown in Figure 2.5iii for all six samples. Intensity values were used due to their sensitivity in detecting larger particles and multiple populations, which was especially useful at higher temperatures, and in order to aid interpretation of the line widths in the following 1D NMR experiments (Section 2.7.1). Table 2.4 shows the sizes for all six samples in number values from 20–75 °C, which allows for comparison among samples of the size of the most prominent population.

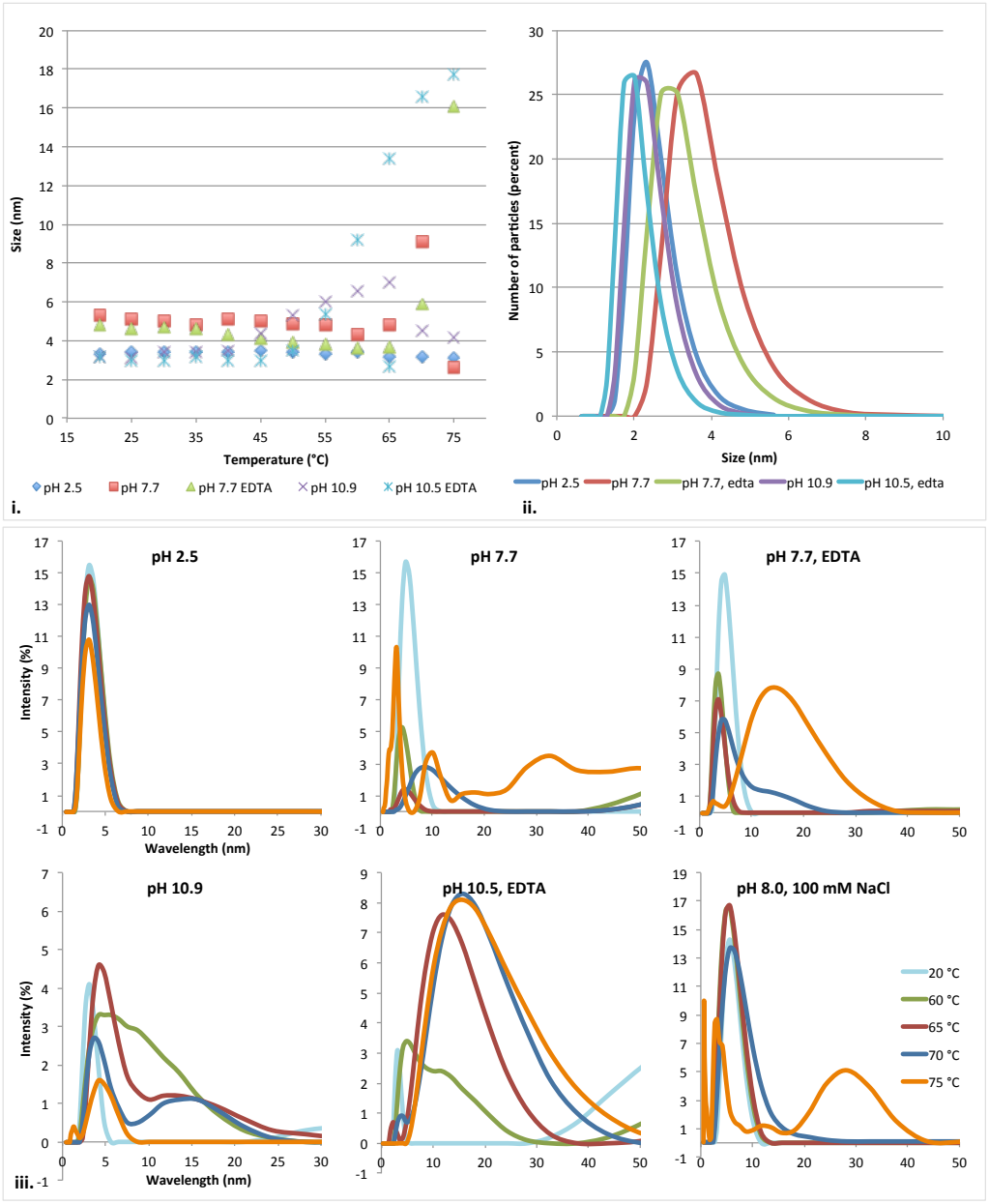


Figure 2.5: Hydrodynamic diameter (nm) of insulin (1 mM) in acidic, neutral (EDTA-free, EDTA-containing) and basic (EDTA-free, EDTA-containing) pH in 10 mM sodium phosphate buffer solutions shown by DLS. i. Dynamic light scattering temperature experiment. The size of insulin particles in terms of intensity values (nm) is shown. ii. Size distribution of the five samples at 25 °C. The number of particles (percent) of each population is shown. iii. Size distributions for all six insulin samples at 20 °C and from 60–75 °C. The particle sizes are shown in intensity values.

Table 2.4: DLS temperature experiment showing the size (hydrodynamic diameter) of insulin particles in terms of number values (in nm). Insulin solutions are 1 mM at acidic, neutral (EDTA-free, EDTA-containing) and basic (EDTA-free, EDTA-containing) pH in 10 mM sodium phosphate buffer and 1.4 mM protein solution of pH 8.0 in 100 mM NaCl, 10 mM sodium phosphate buffer. The percentages, where mentioned, describe the population of each size in the sample. 100% population is implied where no percentages are mentioned.

	20 °C	25 °C	30 °C	35 °C	40 °C	45 °C	50 °C	55 °C	60 °C	65 °C	70 °C	75 °C
pH 2.5	2.3	2.4	2.3	2.3	2.3	2.4	2.1	2.0	2.1	2.1	2.0	2.2
pH 7.7	3.7	3.7	3.4	3.4	3.5	3.4	3.2	3.4	3.2	3.7 (87%) 1.7 (13%)	5.3	2.6 (50%) 8.7 (12%)
pH 7.7 EDTA	3.4	3.2	3.2	3.3	3.0	2.9	2.9	2.7	2.6	2.6	3.3	2.5 (62%) 8.5 (38%)
pH 10.9	2.5	2.3	2.8	2.6	2.5 (73%) 1.2 (27%)	2.9	3.4 (74%) 1.6 (26%)	3.1 (59%) 1.6 (41%)	2.4	2.8	2.7	2.8 (51%) 1.1 (49%)
pH 10.5 EDTA	2.8	2.3	2.4	2.4	2.4	2.4	2.6	2.5	3.3 (83%) 0.9 (17%)	2.2 (83%) 6.4 (17%)	5.1	2.0 (14%) 3.1 (19%) 7.6 (67%)
pH 8.0, 100 mM NaCl	4.1	4.0	3.7	3.8	3.8	3.7	3.7	3.6	3.4	3.5	3.7	2.9 (33%)

2.5. CD Spectroscopy and Secondary Structure Analysis

The CD spectra for different insulin samples at 20 °C (Figure 2.6) all have obvious α-helical content, but differ somewhat from the characteristic α-helix spectrum[90] with 1–2 nm shifts to longer wavelength compared to the typical bands at 195 nm/208 nm/222 nm of an all α-helical protein. These shifts reflect the relatively high β-sheet content of insulin, as the classic β-sheet spectrum has a positive peak between 195 nm and 202 nm and a single negative signal between 215 nm and 219 nm. The relative intensities of the peaks are consistent with the previously reported pH dependence of the insulin CD spectra.[91] The most significant variation concerns the 222 nm band, which increases in magnitude according to: basic pH < acidic pH, neutral-EDTA < neutral EDTA-free (Zn available) sample following the change from predominance of monomer, to dimer, to hexamer.[10] The largely monomeric basic samples present the lowest intensities of the ~195 nm positive peak and the ~222 nm negative peak in accord with an increase in random coil content (usually a negative signal at 200 nm and little, perhaps positive intensity at 222 nm)[92] consistent with the increased flexibility of the N- and C-termini of the B chain of the monomer.[57,93] In the neutral EDTA-free sample, the 222:208 nm ratio is almost 1:1 which is generally accepted to reflect increased hydrophobicity of the environment for the respective helices as would occur in the hexamer where surfaces are protected from the polar aqueous environment.[90] The ~195 nm peak increases are a result of the stabilisation of β-sheets between chains in the oligomeric complexes.

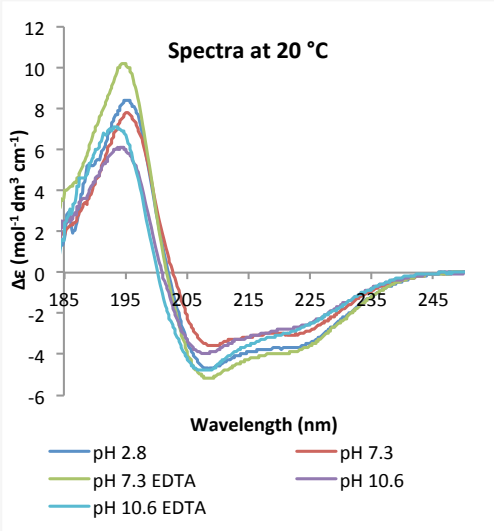


Figure 2.6: CD spectra in Δε units for insulin at 0.1 mg/mL, 10 mM sodium phosphate buffer at the indicated pH conditions at 20 °C.

The differences between the samples were more apparent when we measured the CD as a function of temperature. As described in Section 1.2, thermal denaturation leads to activation of the energy barriers and to high-energy unfolded states of the protein due to loss of the native interactions. In the effort of the protein to reach an alternative conformation, located in one of the neighbouring minima in the energy funnel, multiple off-pathway events can take place resulting in partially unfolded, misfolded structures and aggregated forms. All samples showed a gradual reduction in α-helical character (222 and 208 nm signals, Figure 2.7i), but the temperature at which this occurred and the degree of shift towards the 200 nm negative band of the random coil differed. The negative maximum for each one of the samples at 110 °C (bold line) is shown in Figure 2.7i: the basic samples presented the lowest wavelength (~201–202 nm), perhaps indicating that they are closer to the completely unfolded conformation; the neutral samples shifted down to ~204–205 nm, perhaps due to the presence of more than one oligomerisation state, while the acidic sample presented the highest wavelength (207.6 nm), probably suggesting that a partially unfolded conformation was reached. We quantified these changes using SSNN[70,94] (a neural network algorithm which uses a self-organising map for protein secondary structure assignment from CD spectra and is based on reference set 7 protein database[92] enhanced by 5 denatured and helical proteins) and SELCON3 using the same reference set 7 (Figure 2.8).[71,86] Reference set 7 has been annotated by DSSP. All five samples showed similar, but not identical, trends (Figure 2.7ii), namely gradual decrease in their helical content (of the order of 20–40%), increase in unordered content (15–25%), and a relatively abrupt increase in β-sheet content (10–20%) at different temperatures. In general, the data reflect the higher stability of the oligomers relative to the monomers. The *formation* of β-structure with temperature increase is intriguing and discussed below in the context of the molecular modelling.

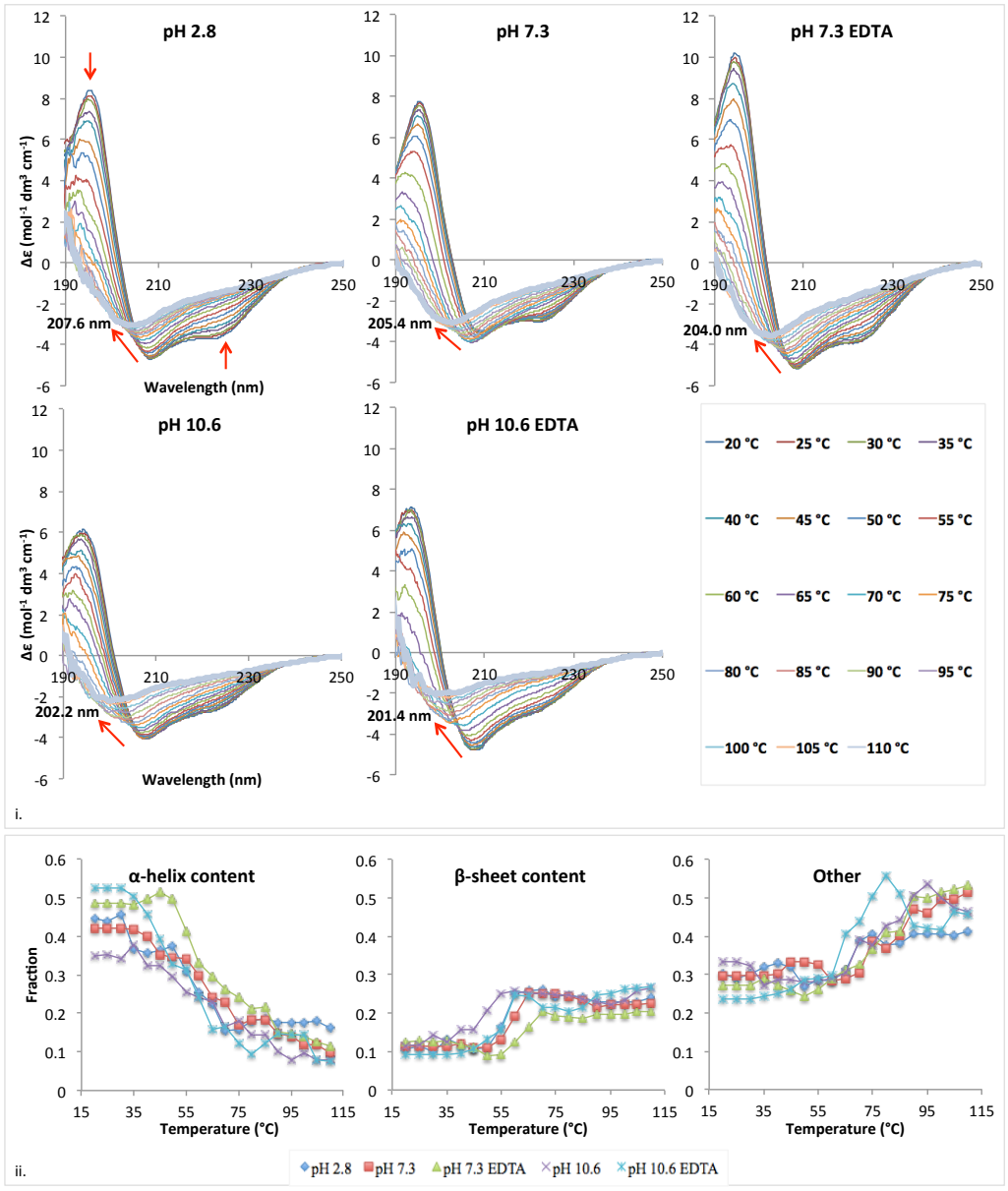


Figure 2.7: i. Changes in the CD spectra of insulin with increasing temperature. The red arrows in the plot of the acidic sample (pH 2.8) show the direction of the decreasing signal intensities at 195 nm, 208 nm and 222 nm with increasing temperature, which was observed for all five samples. For each sample, the spectrum at 110 °C is shown as a bold line, and the wavelength of the negative maximum is given. ii. Changes in secondary structure content (fraction values) of the five different samples of insulin with increasing temperature. The CD spectra were analysed with SSNN.

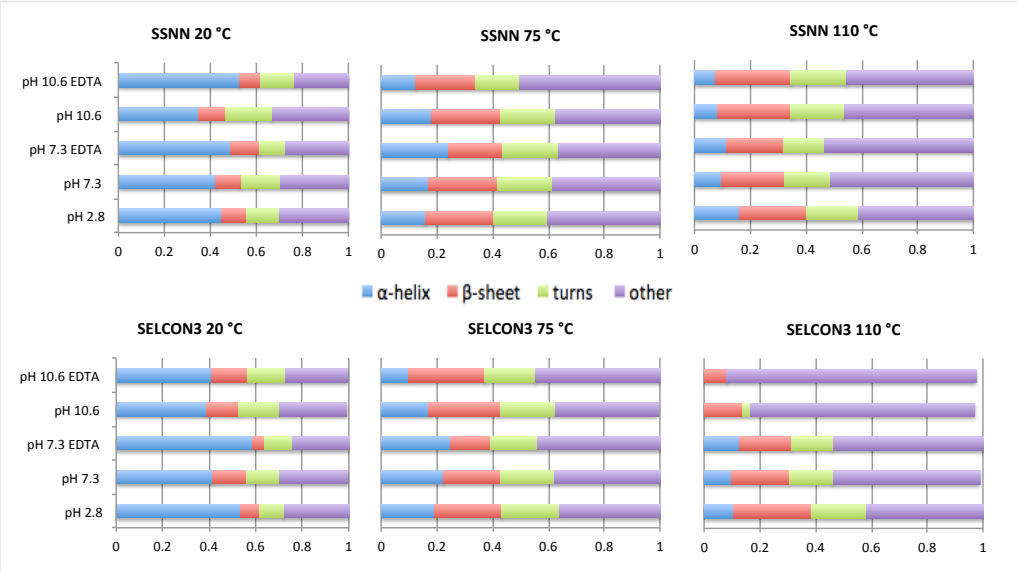


Figure 2.8: Secondary structure content of the five different samples of insulin at 20 °C, 75 °C and 110 °C. The CD spectra were analysed with SSNN and SELCON3. Insulin solutions were 0.1 mg/mL at acidic, neutral (EDTA-free, EDTA-containing) and basic (EDTA-free, EDTA-containing) pH in 10 mM sodium phosphate buffer.

Finally, spectra were recorded at the end of each experiment after cooling down to 20 °C. Partial reversibility was apparent only at acidic pH (Figure 2.9i). The spectra of the neutral and basic samples suggest that irreversible unfolding took place in both cases, due to arrangement of the molecules in trapped intermediate conformations in the folding funnel. None of the samples presented clear signs of aggregation—perhaps due to the low concentration of the samples (~0.1 mg/mL) used for the CD experiments—as visual inspection at the end of the experiment did not show presence of any solid particles or turbidity. Comparison of the HT spectra at 20 °C and after the samples were cooled down to 20 °C (Figure 2.9iii) suggested the appearance of some scattering particles only in the case of the neutral EDTA-free sample at the end of the experiment, as the HT spectrum showed higher values after the sample was cooled down (cooled to 20 °C) compared to the initial spectrum (at 20 °C). The consistent difference in the wavelength of the minima between basic (199.6 nm for both samples) and neutral pH (~203–205 nm) is believed to indicate the presence of the completely unfolded monomer at basic pH, and an equilibrium of partially unfolded hexamers, dimers that have not completely dissociated and unfolded dissociated monomers at neutral pH[95] (also discussed in the molecular modelling, Section 2.8). The higher wavelength shift in the case of the neutral EDTA-free sample (205.0 nm) is due to the presence of the more stable Zn_2-hexamer as the predominant assembly, which is more resistant to dissociation. In contrast, the EDTA-containing

sample (minimum at 202.8 nm) would dissociate more easily, as it is believed to be a mixture of dimers, tetramers and possibly Zn-free hexamers.

The highly helical spectrum at pH 2.8 after cooling down shows that refolding to a partially folded structure occurred, as the spectrum is very similar to that at 45 °C (Figure 2.9ii). This is in agreement with previous studies suggesting the presence of a common partially unfolded intermediate during fibrillation and random coil generation.[96-98] We assume therefore that this intermediate conformation is shown by the spectrum at 110 °C (with a minimum at 207.6 nm). A possible explanation could be that the dissociation of the dimer (predominant at acidic pH) results in two monomers with different flexibility (discussed in molecular modelling, Section 2.8). Monomer 1 represents the completely unfolded conformation, while monomer 2 is likely to represent the partially unfolded intermediate, which can potentially lead to any of the three pathways, namely refold back to a partially folded state (represented by the 45 °C spectrum here), fibril formation or random coil.

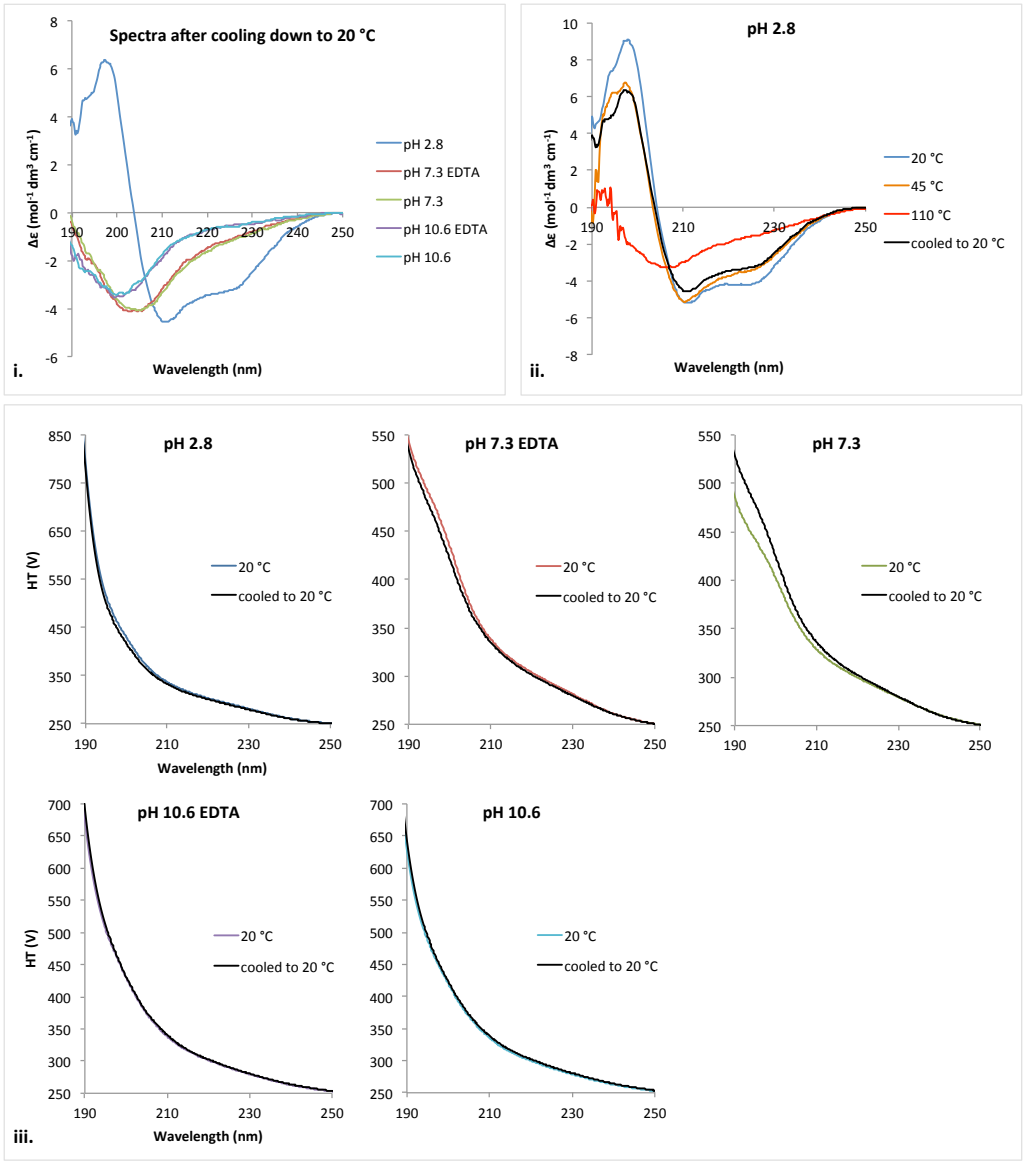


Figure 2.9: i. CD spectra for insulin 0.1 mg/mL at acidic, neutral (EDTA-containing, EDTA-free) and basic (EDTA-containing, EDTA-free) pH in 10 mM sodium phosphate buffer after having been heated and cooled down to 20 °C. Unfolding of the protein at neutral and basic pH seems to be irreversible, while refolding to a partially folded conformation is apparent in acidic conditions. ii. Comparison of CD spectra for insulin at pH 2.8 at 20 °C, 45 °C, 110 °C and after being cooled down to 20 °C. Upon cooling, the partially unfolded intermediate appearing at 110 °C seems to refold to a structure very close to that observed at 45 °C. iii. Comparison of the HT spectra of the five insulin samples at 20 °C and at the end of the experiment, after the samples were cooled down to 20 °C.

Figure 2.10 shows the CD spectra for all five samples at the temperature where the partially unfolded intermediate is observed for the first time. As described above and as expected from the stability order, the intermediate structure is first observed at 65 °C for the two basic samples, at 80 °C for the acidic sample and at 90 °C for both neutral samples.

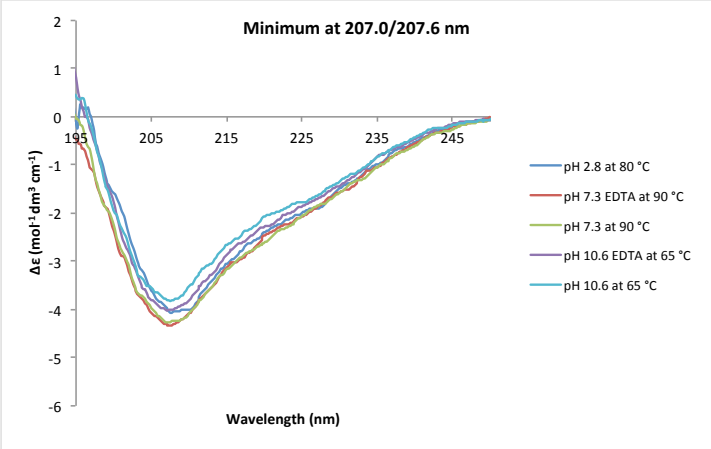


Figure 2.10: CD spectra for each one of the five samples at the temperature where they first present a minimum at 207.0 or 207.6 nm. This structure is believed to represent that of the partially unfolded intermediate and is observed at an earlier temperature for the two basic samples (65 °C), at 80 °C for the acidic sample and at 90 °C for the two neutral samples, following their expected stability order as mentioned in the text.

CD uses a significantly lower protein concentration in comparison to DLS and NMR. In order to match the DLS and NMR experiments, CD spectra of the EDTA-free samples were obtained at a concentration of ~1 mM and compared with the respective data at the usual concentration of 0.1 mg/mL (~17 μM). The comparison is shown in Figure 2.11. Data below 200 nm were excluded as too noisy due to high HT (High Tension) values.[99,100] In all three cases (acidic, neutral and basic pH), higher concentration results in increase of the intensity of the negative peak at 222 nm, which as mentioned above is a sign of increased hydrophobicity. At acidic pH, the dimeric unit is probably moving towards the formation of tetramers but with the dimer dominant as shown by the DLS results. At neutral pH the two spectra are almost identical with the tetramer/hexamer equilibrium slightly moving towards the hexamer formation at the higher concentration, while at basic pH the predominant monomer at lower concentration seems to move towards a monomer/dimer equilibrium at higher concentration (see also DLS and ^{1}H-NMR results). The above data are therefore confirming that the oligomerisation equibria in the case of insulin are concentration-dependent, as expected.

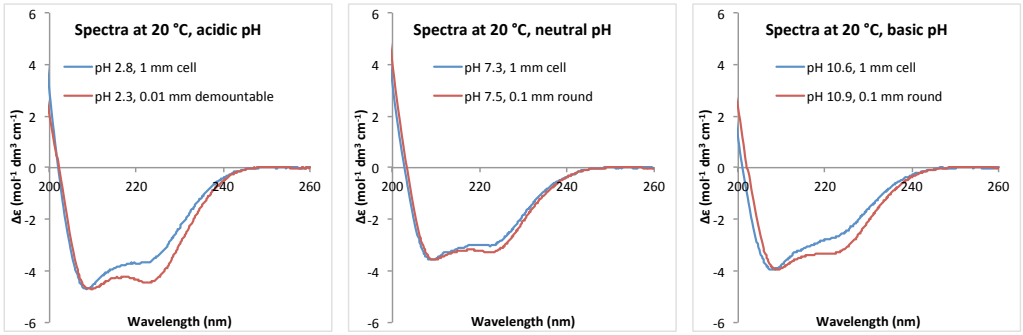


Figure 2.11: CD spectra of EDTA-free insulin samples at acidic (left), neutral (centre) and basic (right) pH. The used concentration of ~1 mM (in red) is in accordance with the concentration used for the DLS and NMR experiments and the spectra are compared with their respective at a concentration of ~17 µM (0.1 mg/mL, in blue).

2.6. FT-IR Spectroscopy and Secondary Structure Analysis

Fourier transform infrared spectroscopy (FT-IR) has been largely used to track changes in the secondary structure of proteins as a function of solvent, pH, concentration and temperature. In the case of insulin, numerous studies have focused on monitoring the aggregation and fibrillation process by following characteristic changes on the amide I band, as usually there is a difference in the band shape between native and thermally denatured states. The majority of the insulin aggregation studies in the literature have been conducted at acidic pH (conditions that promote dimers and monomers), and fibril formation was initiated by heating at 65 °C for several hours.[19-24]

In protein FT-IR spectroscopy, the use of heavy water (D_2O) is very common in order to minimise the H_2O bending vibration band at 1645 cm^{-1}, which overlaps with the amide I vibration band of the protein, and spectra are often acquired using a CaF_2 transmission window. Analysis of the secondary structure content is mainly based on band fitting[19] or the use of neural networks.[101-103] In brief, band fitting involves deconvolution of the amide I peak into specific components corresponding to secondary structure elements. Fourier deconvolution or second derivative analysis[60,104] is used for enhancement of peak positioning, and the peaks are fitted with a Gaussian or Lorentzian function. Integration of the resulting peaks gives the relative amounts of each secondary structure element. Neural networks (NN) are algorithms trained to infer the secondary structure of proteins (pattern recognition). The training involves a reference set containing protein spectra with known secondary structures, and the analysis (fitting) is largely affected by the band shape of the tested peak.

In this experiment insulin was used at higher pH (8.2) than usual, under the hexamer promoting conditions,[25] which is believed to protect insulin from fibrillation.[98] Water-based buffer was preferred, as this represents better the protein's native environment in comparison to D_2O, and the ATR technique was used. Our neural network (SSNN) was retrained[70] with an FT-IR reference set consisting of 47 protein spectra recorded in the transmission mode. Comparing ATR to transmission spectra might affect the analysis somewhat due to the effect of the anomalous dispersion, which is induced by the use of ATR and results in shifting the peak positions (see Section 1.3.2 for FT-IR). The shape of the band though is not expected to change much, thanks to the large amount of protein that was deposited on the crystal (4 solid layers of the 10 mg/mL protein solution). The depth of penetration is also an

effect specific to ATR, which changes the relative peak heights. However, since we only use the normalised amide I peak, our analysis will not be affected by this effect.

In the present study, insulin fibrils were obtained as described in materials and methods (Section 2.2.3), and a comparison between insulin as powder from the purchased bottle and the formed fibrils is shown in Figure 2.12. Figure 2.13 shows the ATR FT-IR spectra of insulin at 25 °C, after being heated at 75 °C, 95 °C and of insulin fibrils in absolute absorbance units (i) and normalised absorbance (ii). As the temperature increases the amide I peak becomes wider (Figure 2.13 and Figure 2.14) which is typical of unfolded structures,[22] while also a decrease in intensity is observed, perhaps denoting a decrease in concentration in the supernatant solution which was used for the spectra (as the protein precipitated in the form of aggregates or fibrils). As a consequence of the reduced concentration, the amide I/amide II peak ratio is affected by the penetration depth effect at higher temperatures, and is therefore distorted (Figure 2.13, the two peaks are almost of equal height at 95 °C and the fibres spectrum).

Figure 2.12: Aluminium weighing boats containing insulin from the purchased bottle as white powder (left), and the fibrils that were left at room temperature for five months as pale yellow solid (right) are shown.

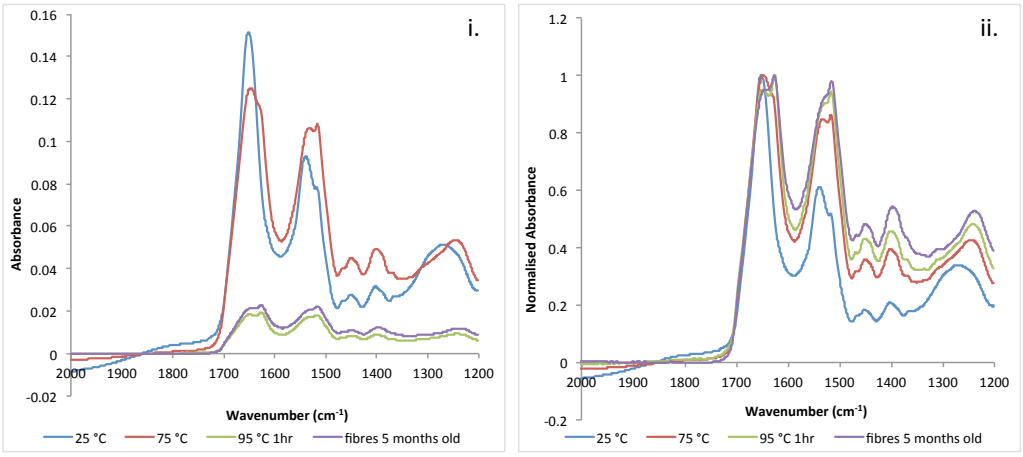
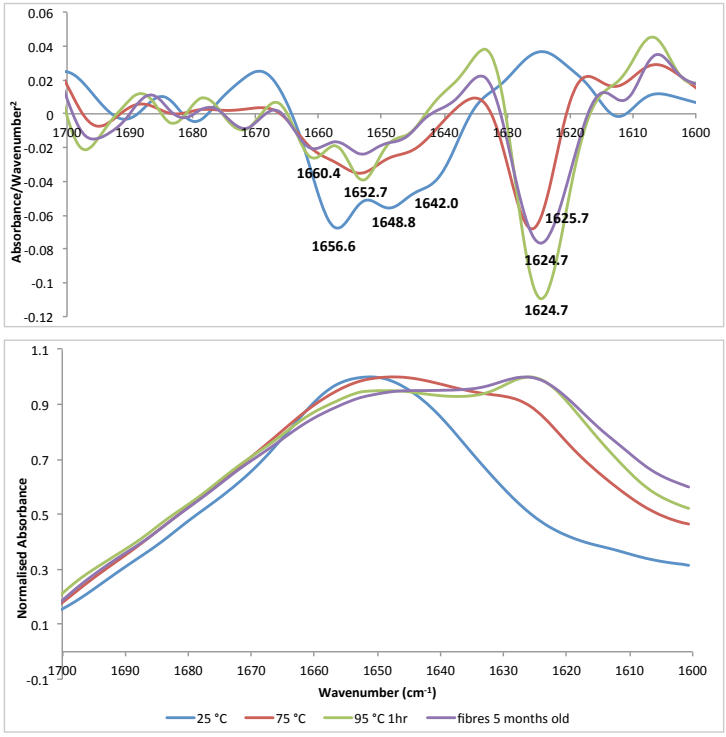


Figure 2.13: ATR-FTIR spectra of insulin at 25 °C, after being heated at 75 °C, 95 °C and of insulin fibres after being left at room temperature for five months. i. The spectra are shown in absolute absorbance units. ii. The amide I band is normalised to unity.

At 25 °C the amide I peak appears at 1650.8 cm^{-1} while resolving the peak with the second derivative (Figure 2.14) reveals its positions at 1656.6 cm^{-1} (characteristic band frequency of proteins with high α-helix content), 1648.8 cm^{-1} (usually referred to as random coil) and a shoulder at ~1642.0 cm^{-1} (β-sheet in native proteins).[105,106] Heating the sample at 75 °C results in shifting the amide I peak to lower frequency (1647.9 cm^{-1}) evidencing an increase in the random coil content, while at the same time a shoulder appears at ~1627 cm^{-1} which is typical for the formation of parallel β-sheets resulting from aggregation of unfolded proteins. The second derivative locates a wide peak at 1652.7 cm^{-1} with a shoulder at ~1642 cm^{-1} and a peak at 1625.7 cm^{-1} (fibril formation). The 95 °C and the fibrils spectra are quite similar, presenting the amide I peak at 1647.9 cm^{-1} and a prominent feature at 1625.7 cm^{-1} (the observed shoulder at the 75 °C spectrum). The second derivative resolves the peak giving maxima at 1660.4 cm^{-1} (turns or 3_{10} helix), 1652.7 cm^{-1} (disordered) and 1624.7 cm^{-1} (parallel β-sheet in fibrils).



Band position (cm^{-1})	Secondary structure
1660	turns or 3_{10} helix
1657	α-helix
1653	α-helix or random coil
1648	random coil
1642	β-sheet in native proteins
1626	β-sheet in fibrils

Figure 2.14: The amide I peaks for the spectra at 25 °C, 75 °C, 95 °C and for insulin fibrils are normalised to unity (bottom). Second derivative amide I spectra (top) are used to aid peak positioning. The resolved band positions (within ±1 cm^{-1}) are given in the table (top right). The respective secondary structures were assigned according to Kong *et al.*[107]

Quantification of the secondary structure elements using SSNN showed a total decrease of 22% in the α-helix content and 17% increase in the β-sheet content for the fibres spectrum (Table 2.5). Although the assignment is good enough for helices and sheets, random coil content does not seem to be affected which is clearly in disagreement with what happens in reality. As explained in Section 1.2, thermal experiments can lead to a variety of off-pathway events in the folding funnel (high-energy states), including non-native conformations of the protein (thermal unfolding), trapped intermediates between the denatured and the native state, and aggregation of misfolded structures. In the case of insulin, aggregation is initiated by a partially unfolded intermediate of the monomer,[21,97] and therefore transition from the hexamer to the monomer results in loss of ordered structure due to the loose termini, as does partial unfolding of the monomer at high temperatures prior to aggregation.

Table 2.5: Secondary structure assignment for the ATR FT-IR spectra of insulin at 25 °C, 75 °C, 95 °C and for insulin fibrils using the IR trained version of SSNN.

Sample	α-helix	β-sheet	bend	turn	coil
25°C	0.47	0.10	0.08	0.12	0.23
75°C	0.33	0.18	0.11	0.15	0.24
95°C 1hr	0.32	0.18	0.11	0.15	0.24
fibres 5 months old	0.25	0.27	0.11	0.14	0.23

By inspecting outputs from the analysis of the FT-IR spectra with SSNN, a poor spectral fit and an increased error (NRMSD value, normalised root mean squared deviation) were observed at higher temperatures. Figure 2.15 shows a comparison of the original and the fitted (predicted) spectra, and the resulting residual. Indeed SSNN seems to fail in recognising the new feature at 1625.7 cm^{-1} and probably this is due to the absence of any spectra of denatured proteins forming aggregates or fibrils from the reference set. Therefore the occurring change in the shape of the amide I band (an additional maximum between 1620–1640 cm^{-1}, characteristic of intermolecular β-sheet resulting from fibril formation of unfolded proteins) is not represented in the reference set. Hence, caution has to be exercised when interpreting the results, and the use of SSNN in the estimation of secondary structure from ATR FT-IR spectra should be regarded as an approximation, as different conditions (*e.g.* comparing ATR to transmission spectra) make precise assignment of the secondary structure and frequency challenging.

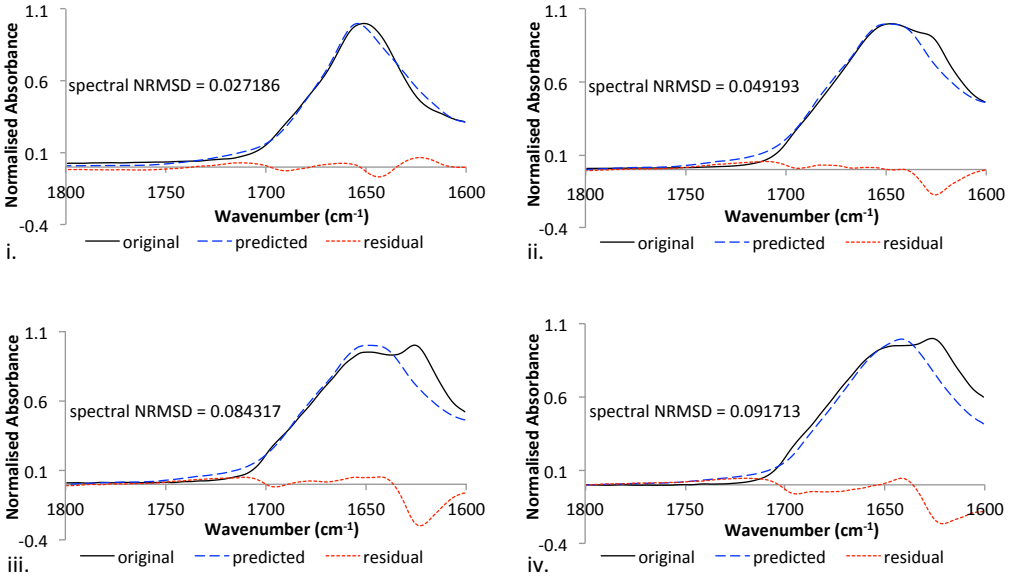


Figure 2.15: SSNN output from the analysis of the ATR FT-IR spectra at 25 °C (i), 75 °C (ii), 95 °C (iii) and for insulin fibrils (iv). The spectral fit of the original (black) to the predicted spectrum (blue dashed line), their residual (red) and the NRMSD value are shown in each case.

2.7. 1D ^{1}H- AND 2D-NMR SPECTROSCOPY

2.7.1. 1D ^{1}H-NMR results

Applications of NMR spectroscopy to study protein structure and dynamics in solution typically involve an impressive arsenal of state-of-the-art multi-nuclear multi-dimensional techniques such as nuclear Overhauser effect spectroscopy (NOESY), hetero-correlated (^{1}H, ^{13}C, ^{15}N) experiments and their hyphenated equivalents, ^{15}N-NMR-based relaxation measurements to study protein dynamics, and diffusion-ordered spectroscopy (DOSY) for estimating molecular size and oligomer speciation.[108] In the case of insulin especially, specific ^{1}H-NMR experiments have been used for detection of aggregation.[109,110] However, as our goal was to detect (as simply as possible) qualitative differences between two samples, we chose to exploit the information-rich,[111] but often hard to interpret, 1-dimensional ^{1}H-NMR spectroscopy as a simple and relatively fast approach. The time required to record each spectrum did not exceed 5 mins and approximately 4 hours were needed in total for the temperature-dependent measurements. Key diagnostic variations were observed in line widths (measured values are given where possible), chemical shifts and peak dispersion as a function of solution conditions.

Line widths and peak dispersion (Figure 2.16): The *aromatic and backbone* H^{N} signals between 6 and 10 ppm show clear differences in line widths (Figure 2.16i) between samples at room temperature. At *basic pH*, most of the backbone H^{N} signals have disappeared due to exchange with the water protons on an intermediate timescale, and the narrow line widths of the signals for aromatic and aliphatic protons suggest that mono- and dimeric states are favoured.[112] The broadening of the aromatic peaks above 40 °C (Figure 2.16ii C, D) is probably an indication of the onset of aggregation, which is also supported by the appearance of larger particles in the DLS data especially at temperatures from 60 °C onwards (Figure 2.5iii).

At *acidic pH*, the backbone H^{N} peaks are widely dispersed (Figure 2.16i), indicating that the protein is in a well-folded state, and sharp enough to confirm that smaller oligomers, such as dimers, are predominant in these conditions.[48]

The broader peaks throughout the *neutral-EDTA* spectra (up to 55 °C, Figure 2.16ii B) are a sign of shorter relaxation times, which are, amongst other parameters, the result of the presence of larger oligomers (*i.e.* tetramers and hexamers) with longer rotational correlation times (τ_c). Above 55 °C the decrease in line width of the aromatic peaks around 6.4–7.6 ppm (Figure 2.16ii B) can either result from faster tumbling of

the present molecules at higher temperatures or indicate that dissociation to smaller oligomers (*e.g.* dimers) occurs. Aggregation seems to start at 70 °C as the lines become broad again. As shown above, in the DLS data (Section 2.4, Figure 2.5iii and Table 2.4), the neutral EDTA-containing sample presented particles that decreased in size with increasing temperature. Particles with a diameter of 2.6–2.7 nm (number values in Table 2.4, close to the size of monomers and dimers that were observed in the basic samples) were detected at 55–65 °C, and 2.5 nm and 8.5 nm (number values) at 75 °C, further supporitng the case of dissociation of larger to smaller oligomers from 55 °C onwards, and onset of aggregation at 70 °C (as can be seen in Figure 2.5iii).

The pronounced broadness of all peaks in the spectrum of the *neutral EDTA-free* sample is indicative of the dominance of the hexamer[12] (Figure 2.16ii A), while at higher temperatures (60–70 °C) the aromatic peaks (6.4–7.6 ppm) become progressively sharper. Similarly to what was described for the neutral EDTA-containing sample, sharper lines can either result from faster tumbling of the present molecules or suggest an increase in the presence of dimeric or monomeric units.[34] In this case, DLS data (Figure 2.5iii and Table 2.4) were less clear. However, a decrease in the particle diameter was observed at 60 °C (3.2 nm, number values as presented in Table 2.4), which probably indicates an equilibrium of larger and smaller oligomers, while smaller particles (2.6 nm, close to the size of monomers and dimers) were detected at 75 °C, as well as multiple populations of larger particles (Figure 2.5iii). The additional hexamer control sample, at pH 8.0, presented similar behaviour in DLS to the neutral EDTA-free sample, namely gradual decrease in particle size with increasing temperature and smaller particles (2.9 nm, Table 2.4) as well as larger particles at 75 °C (Figure 2.5iii). The line widths for the peak at 6.71 ppm were measured over the increasing temperature (Figure 2.16ii A) to give an example of how this feature can be exploited in a quantitative manner for quality control of proteins where molecular size or oligomerisation state is relevant, and to complement DLS data.

Aromatic region (6.4–7.6 ppm) (Figure 2.16ii): Insulin has Tyr and Phe residues, which have protons that can be monitored to report on their environment. Qualitatively, it is evident that the two *basic samples* have a high degree of similarity in this region, whilst the *two neutral samples* are significantly different. The Tyr (B26) Hδ proton signal at 6.63 ppm, characteristic for the monomeric state,[15] shows a gradual decrease and eventual disappearance at 50 °C for the two *basic samples* (Figure 2.16ii C, D) indicating the transition from the monomeric to an aggregated state.

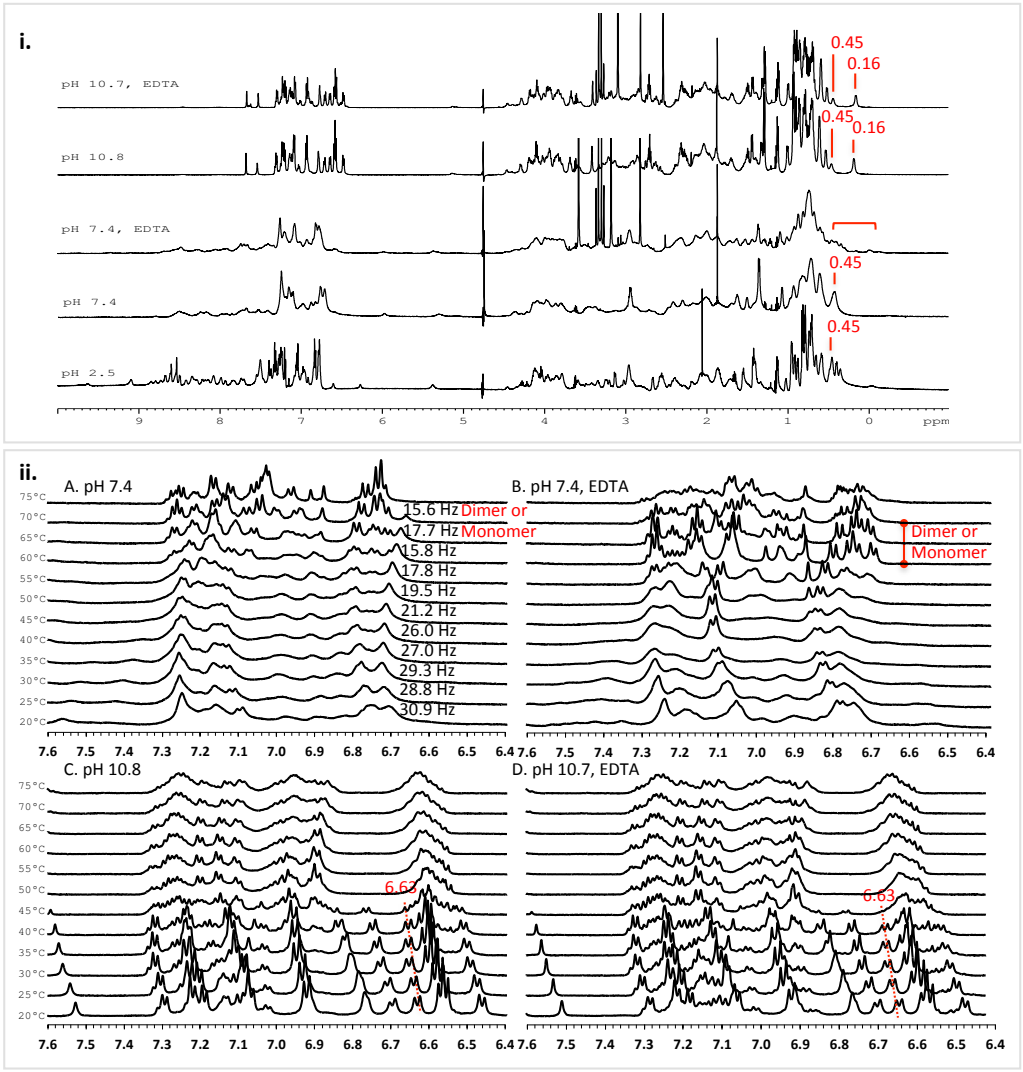


Figure 2.16: i. Comparison of line widths and overall dispersion in the ^{1}H-NMR spectra of insulin (1 mM) in acidic, neutral (EDTA-free, EDTA-containing) and basic (EDTA-free, EDTA-containing) pH, 10 mM sodium phosphate buffer solutions at 25 °C. In the aliphatic region, at 0.16 and 0.45 ppm the characteristic peaks for the monomer and the dimeric unit respectively are indicated. ii. Aromatic region (6.4–7.6 ppm) of the ^{1}H-NMR spectra of insulin at neutral and basic pH at increasing temperature. The sharpness of the peaks at 70 °C for the pH 7.4 EDTA-free sample (tetramers and hexamers predominant at room temperature) (A), and at 60–65 °C for the pH 7.4 EDTA-containing sample (tetramers predominant at room temperature) (B), suggests an increase in the presence of dimeric or monomeric units. The line widths for the peak at 6.71 ppm for the neutral EDTA-free sample (A) were measured at half height of the peak, indicating transition towards smaller oligomers with the increase of temperature. The peak at 6.63 ppm up to 45 °C for both basic samples pH 10.8 EDTA-free sample (C) and the pH 10.7 EDTA-containing sample (D) (monomers and dimers predominant at room temperature) refers to a Tyr (B26) Hδ proton signal characteristic for the monomeric state.

Histidine Hε1 region (7.45–8.70 ppm) chemical shifts and line widths (Figure 2.17i): Histidine NMR signals are particularly sensitive to pH as their imidazole rings gain/lose a proton between pH 6–7, unless they are buried, hydrogen-bonded, or metal-bound (Figure 2.17i B). *Basic samples* (Figure 2.17i D, E) have two strong signals at 7.51 and 7.66 ppm at room temperature, with measured line widths 4–6 Hz (representing the monomer), which have been assigned to the Hε1 protons of His

(B5) and His (B10) respectively[14,16] and decrease in intensity when the temperature increases. At 60 °C, a broader peak at 7.61 ppm which has been assigned to a denatured form of insulin replaces these signals. At *acidic pH* (Figure 2.17i A), the histidine Hε1 protons appear at 8.53 ppm for His (B5) and 8.61 ppm for His (B10) (line widths 6–7 Hz, representing the dimer), as the imidazole ring is protonated under these conditions.[12,48,113,114] At *neutral pH* (Figure 2.17i B, C), the histidine Hε1 peaks (~7.5–7.6 ppm) appear very broad at 20 °C (34–39 Hz for the EDTA-free sample with the hexamer as the predominant form and 27–31 Hz for the EDTA-containing sample at which the tetramer is favoured), but become progressively sharper as the temperature increases (*e.g.* at 60 °C, the line width is 4 Hz indicating dissociation to monomer). A more detailed discussion about the neutral samples is following. The measured line widths of the histidine peaks give a representation of the oligomerisation state in each sample (Figure 2.17i, Table 2.6).

The presence of the peak at 7.43 ppm in the case of the two *neutral samples* (Figure 2.17i B, C), has been previously reported in the literature to refer to the Hε1 proton of the Zn-free His (B10) signal in higher oligomers (hexamer and half part of the tetramer), where a shielding effect is caused by the histidine rings from the opposite dimeric units.[17] Alhough both of these conformations (tetramers and possibly Zn-free hexamers due to their high association constant)[45] are favoured in the case of the EDTA-containing sample (complexed Zn), in the presence of Zn (EDTA-free sample) the formation of the Zn_2-hexamer is preferred to the Zn-free molecule,[115] thus the 7.43 ppm peak is believed to represent only the Zn-free tetramer (half part). In the other half of each tetramer and in dimers, the same proton of His (B10) is exposed and deshielded at 7.71 ppm as previously reported.[17] The broad peak at 7.58 ppm, more obvious in the case of the neutral sample containing Zn (EDTA-free), is believed to represent both the His (B5) Hε1 proton—in the hexamer due to its upfield shift—and the Zn-coordinated His (B10) resonance in the hexamer where the downfield effect from the Zn-binding is balanced by the upfield-shifting effect from the opposite rings.[12,14,17] The peak at 7.66 ppm in the EDTA-containing neutral sample represents the His (B5) Hε1 proton of smaller oligomers such as dimers and tetramers. Therefore, the above evidence suggests the presence of tetramers and hexamers as predominant species at neutral pH in the presence of Zn (EDTA-free sample) and the presence of dimers, tetramers and probably Zn-free hexamers in the case of the EDTA-containing neutral sample. The progressive sharpening of the His (B5) and His (B10) Hε1 proton peaks with the increase of temperature is perhaps a sign that the Zn is not bound to

insulin anymore as a consequence of the hexamer dissociation (*e.g.* EDTA-free neutral sample at 70–75 °C, Figure 2.17i B), as seems to be the case in acidic conditions (Figure 2.17i A) in which His (B10) is unable to coordinate metal ions.

The sharpening of the Histidine Hε1 peaks was followed with the increase of temperature for all five samples and the line widths measured at half height of the peaks (FWHH) are reported in Table 2.6. The FWHH values increase from the monomer (basic pH, ~4 Hz) to the dimer (acidic pH, ~7 Hz) and tetramer/hexamer (neutral EDTA-containing, ~27 Hz/neutral EDTA-free sample, ~35 Hz) and decrease to ~4 Hz at higher temperatures for all samples. The latter observation indicates the transition to smaller oligomers (probably monomers) that subsequently will start aggregating before fibrillation takes place. The line width of the histidine peak that corresponds to the denatured species is also given for the basic samples and those of the characteristic for the dimeric unit and the monomer (at ~0.45 and 0.16 ppm respectively) for the basic samples and the neutral EDTA-free sample, representing the oligomerisation state in each case.

Shifting: At pH 2.5, the upfield shift of a doublet peak from 8.48 ppm (in the 25 °C spectrum) to 8.15 ppm at 65 °C (Figure 2.17i A) is remarkable. This signal is believed to represent the backbone amide proton of Asn (B3)[48] with a coupling constant $^{3}J_{H}{}^{N}{}_{H}{}^{\alpha}$ of 6.7 Hz, and the −0.37 ppm shift can probably be attributed to temperature dependence of the amide proton due to its participation in hydrogen bonding.[116] The flexibility of the N-terminus of chain B, where Asn (B3) belongs, could also support participation of the residue in hydrogen bonding.

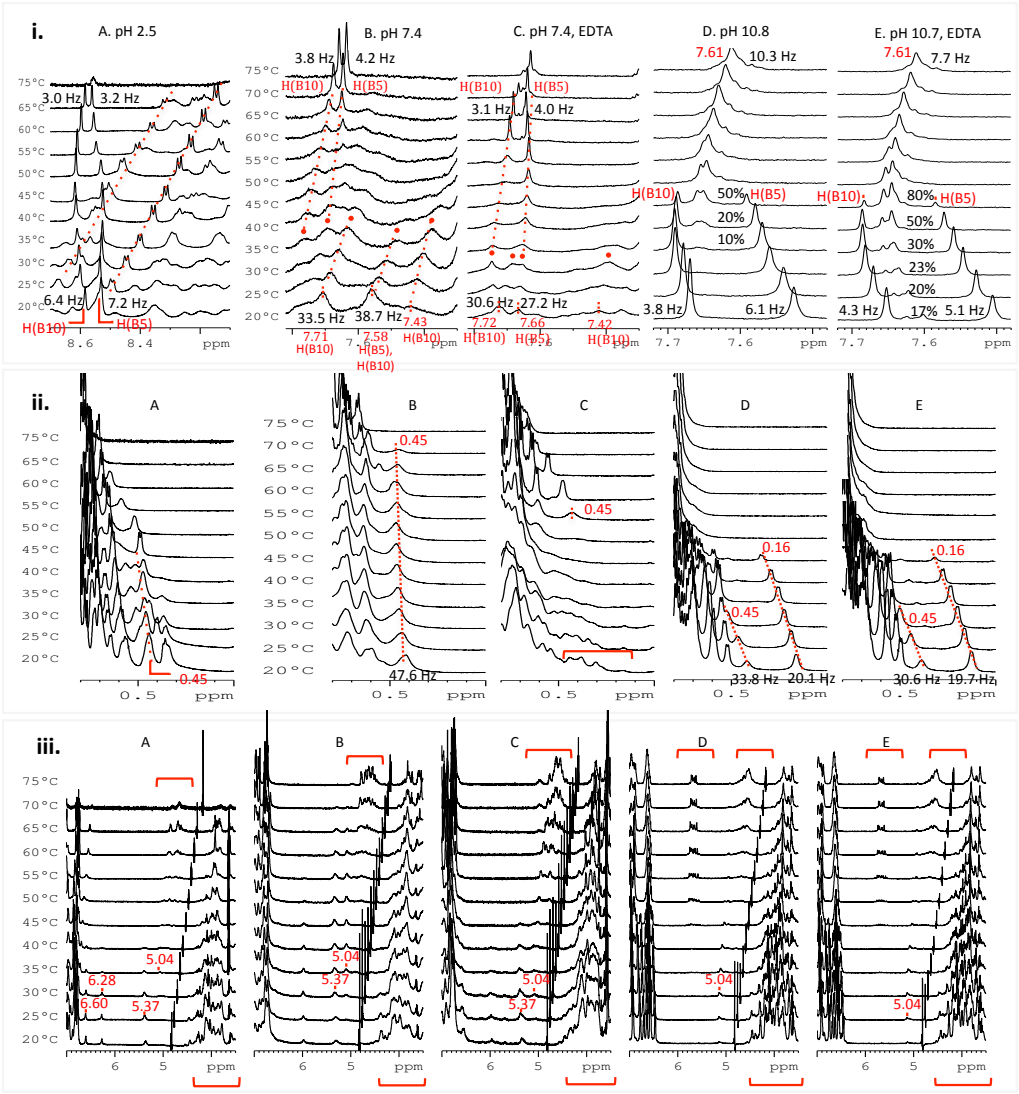


Figure 2.17: Selected regions from the ^{1}H-NMR spectra of insulin at acidic (A), neutral EDTA-free (B) and EDTA-containing (C), basic EDTA-free (D) and EDTA-containing (E) pH at increasing temperature. i. Histidine Hε1 region: A. (8.1–8.7 ppm), B–E. (7.48–7.72 ppm). The His (B5) and His (B10) Hε1 proton signals are shown. The noted line widths of the histidine peaks are an indication of the size of the dominant species in each sample. For the acidic sample the increasing shielding effect on the two indicated doublet peaks (dotted lines; the peak at 8.48 ppm probably corresponds to the backbone H^N of Asn (B3), while the signal at 8.66 ppm has not been assigned to any residue) is suggested to be a temperature dependence effect due to participation of the protons in hydrogen bonding. For the two basic samples the appearance and increase in the population of a new species, probably corresponding to denatured forms, is indicated. ii. Methyl region (0.0–0.8 ppm). The disappearance of the peaks characteristic of the monomer (0.16 ppm) and dimeric unit (0.45 ppm) is shown, indicating unfolding of the protein with increasing temperature. The shown line widths indicate the presence of small units at basic pH (D, E) and higher oligomers at pH 7.4 EDTA-free sample (B). iii. Hα region (3.5–7.0 ppm). The downfield shift of peaks reveals a gradual increase in the β-sheet content, indicating the formation of aggregates. The peaks at 5.04 ppm and 5.37 ppm belong to Hα protons, while the peaks at 6.28 and 6.60 ppm are side chain NH protons (see 2D-NMR results, Figures 2.19–2.22).

Methyl region (0–1 ppm) (Figure 2.17ii): Insulin has an extremely high-field shifted signal for the Leu (B15) Hδ methyl protons in the monomer that is present at *basic pH*[117-119] (0.16 ppm, Figure 2.17ii D, E). This signal moves downfield for the

dimer and higher assemblies[15,120] (~0.45 ppm at 20 °C) where shielding of this methyl group by Phe (B24) and Tyr (B26) (Figure 2.18) is less effective, though still present.[15] The *acidic pH* and *neutral EDTA-free* samples do not show the resonance at 0.16 ppm, confirming the absence of monomers (Figure 2.17ii A, B), while the addition of *EDTA at neutral pH* led to a high level of heterogeneity in the methyl region, consistent with the presence of several species (Figure 2.17ii C). In all cases, disappearance of all high-field shifted peaks in the methyl region indicates unfolding of the protein. The high stability of the hexamer is thus reflected in the persistence of such high-field shifted signals up to a temperature of 70 °C for the most stable neutral (pH 7.4) EDTA-free sample (Figure 2.17ii B).[34,50] The line widths of these methyl signals are given where possible (Figure 2.17ii, Table 2.6).

Hα resonances (3.5–7.0 ppm) (Figure 2.17iii): For all samples, the downfield shift of peaks in the Hα proton area, from 3.5–4.5 ppm to 4.6–5.8 ppm at increased temperatures (≥50 °C), is thought to reveal gradual increase in the β-sheet content,[121] which is in agreement with our CD and FT-IR results and usually accompanies the formation of aggregates. Based on the knowledge that the chemical shifts of Hα protons are dependent on secondary structure, it has been suggested that integration of 1-D ^{1}H-NMR spectra from 4.85–5.90 ppm (largely Hα resonances from residues in β-conformation) may give estimates of the β-strand content of a protein without the need of prior resonance assignment.[122] Amyloid-like fibrils are composed of arrays of β-strands;[123] it is therefore possible that the increase in low-field shifted Hα proton resonances is an indication for the onset of fibril formation, although the sharpness of the peaks indicates that the species are not highly aggregated. Loss of NMR signal intensity for the acidic sample at 75 °C (Figure 2.17i–iii A) is also a consequence of aggregation which may be observed by visual inspection of the sample at the end of the experiment. Differences in the location of these peaks between the *neutral* and *acidic* (at 4.6–5.0 ppm) (Figure 2.17iii A–C) and the *basic* samples (at 4.6–4.9 ppm and 5.6–5.8 ppm) (Figure 2.17iii D, E) suggest the presence of different misfolded conformations and aggregates in each case, in accord with our DLS data. The peaks with chemical shifts at 5.04, 5.37, 6.28 and 6.60 ppm are discussed below.

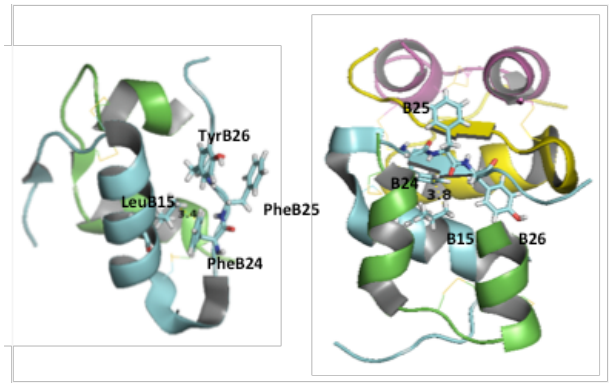


Figure 2.18: The distance between the Leu (B15) Hδ methyl group and the aromatic residue Phe (B24) is shown. *Left*: in the case of the monomer the distance was measured as 3.4 Å. *Right*: for the dimer the distance was measured as 3.8 Å.

Table 2.6: The changes on line widths (in Hz) with temperature are shown. The line widths were measured at half height of the peaks. The selected peaks are: the Histidine Hε1 peaks of the five samples, the Leu (B15) Hδ methyl peaks for the dimeric unit (~0.45 ppm) and the monomer (~0.16 ppm) for the basic samples and the neutral EDTA-free sample (dimeric unit only), and the His peak for the denatured species for the basic samples. The measured line widths are indicative of the oligomerisation state and the predominant species in each sample, namely dimer at pH 2.5, hexamer at pH 7.4 EDTA-free, tetramer at pH 7.4 and presence of EDTA, monomer at basic pH (at 20 °C). Similarly, changes in the line widths with the increase of temperature denote changes in the oligomerisation state.

		Width (Hz) at T (°C)											
Characteristic peaks	**peaks (ppm)**	**20 °C**	**25 °C**	**30 °C**	**35 °C**	**40 °C**	**45 °C**	**50 °C**	**55 °C**	**60 °C**	**65 °C**	**70 °C**	**75 °C**
	pH 2.5												
His (B10)	8.58	6.40	7.97	8.13	9.58	6.13	3.75	3.47	3.16	2.91	3.00		
His (B5)	8.53	7.24	6.61	5.27	4.59	5.85		5.76	5.01	3.95	3.15		
	pH 7.4												
His (B10)	7.71	33.54	33.50	32.03	35.19					20.25	12.99	6.20	3.76
His (B5)	7.56	38.72	42.31	41.59	30.71					17.05	8.51	5.39	4.19
Leu (B15) Hδ methyl (dimeric unit)	0.42	47.60	55.03	59.22	43.93	42.44	42.45	43.72	41.95	41.54	41.19	40.36	
	pH 7.4 EDTA												
His (B10)	7.70							30.65	16.18	4.46	3.12	4.05	
His (B5)	7.66	27.25	26.54	27.92	36.71	34.23	20.10	9.02	4.99	4.01	3.98	3.96	4.36
	pH 10.8												
His (B10)	7.67	3.85	3.88	3.95	3.91	3.51	3.43						
His (denatured species)	7.66						11.36	10.54	10.63	9.58	8.22	8.82	10.29
His (B5)	7.53	6.12	6.39	5.91	5.25	4.09	3.76						
Leu (B15) Hδ methyl (dimeric unit)	0.42	33.75	29.25	19.33									
Leu (B15) Hδ methyl (monomer)	0.16	20.12	19.31	18.06	18.40	15.60	14.03						
	pH 10.7 EDTA												
His (B10)	7.65	4.33	4.01	3.78	3.36	3.22	3.04						
His (denatured species)	7.64									7.28	6.65	6.93	7.67
His (B5)	7.51	5.12	4.24	3.88	3.52	4.48	3.60						
Leu (B15) Hδ methyl (dimeric unit)	0.39	30.65	25.07	20.41									
Leu (B15) Hδ methyl (monomer)	0.13	19.73	19.00	17.91	17.61	15.81	16.36						

2.7.2. 2D-NMR results

Interesting is the fact that different peaks appear for each sample in the region from 5.0–6.6 ppm (Figure 2.17iii), with the only common peak for all samples being the one at 5.04 ppm. In an attempt to determine the identity of the residue to which this proton refers, 2D ^{1}H experiments were carried out. The 2D-TOCSY spectrum at pH 2.5 at 25 °C (Figure 2.19) showed a correlation of this resonance with resonances at 8.53 ppm and 2.97 ppm. As discussed above, the Hε1 proton of the protonated His (B5) residue appears at 8.53 ppm, therefore the 5.04 ppm is likely to represent the Hα proton of His (B5), whilst the resonance at 2.97 ppm may refer to the Hβ protons.

The peak at 5.37 ppm is also a Hα proton and presents correlations with two spin systems, one of which probably belongs to an aromatic residue as it correlates with two aliphatic peaks at 2.81 and 3.35 ppm and an aromatic peak at 7.22 ppm (Figure 2.19).[16] This peak is tentatively assigned to Phe (B24) on the C-terminus of chain B on the monomer-monomer interface, due to its further NOESY correlations (Figure 2.20, in green). The two neutral samples present peaks at 5.04 and 5.37 ppm, which have also been reported in the literature for the ^{1}H-NMR spectrum of the native Zn_2-hexamer in the Zn(II)-T_6 state,[43,124,125] while in the two basic samples only the peak at 5.04 ppm appears.

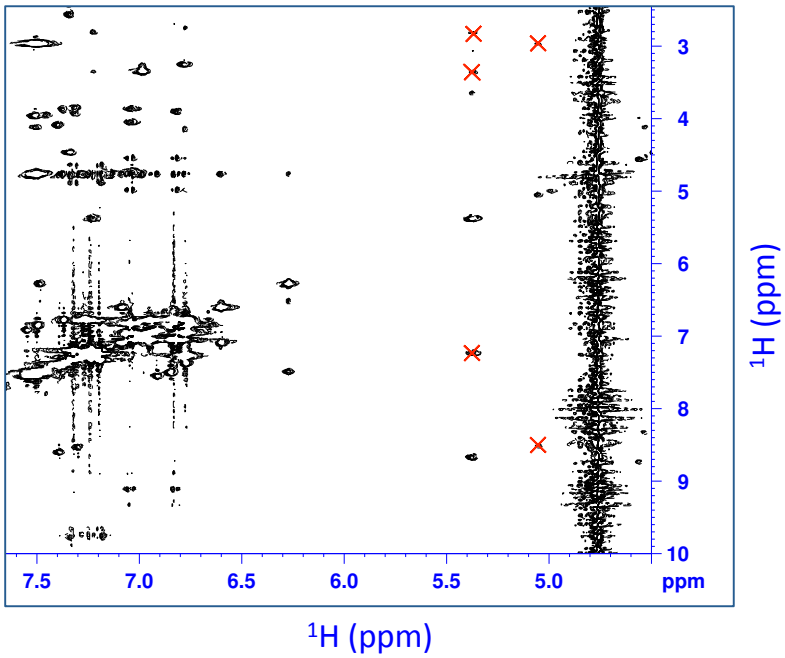


Figure 2.19: A portion of a 2D-TOCSY spectrum at pH 2.5 25 °C, showing the correlations of the protons at 5.04 and 5.37 ppm.

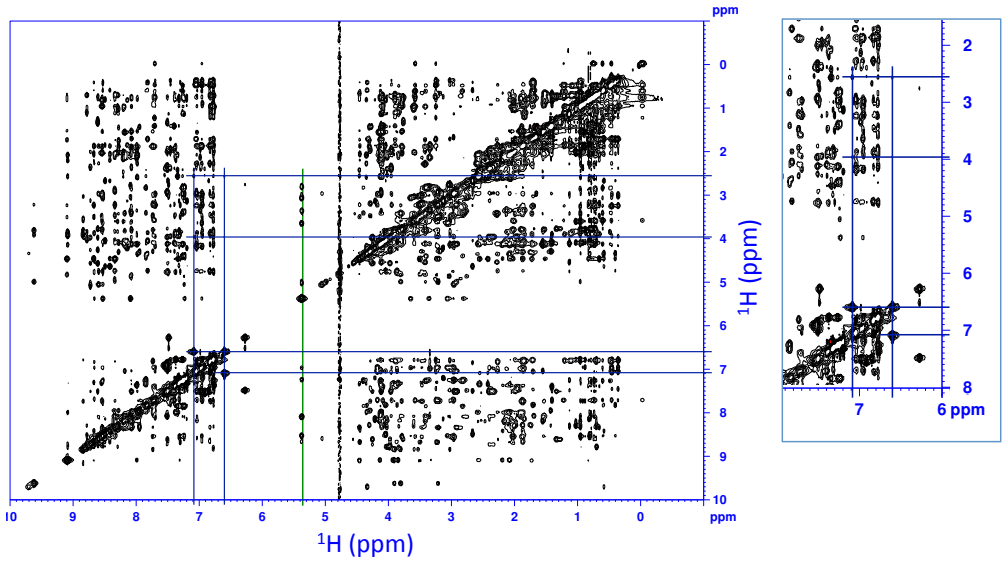


Figure 2.20: ***Left*****: 2D-NOESY spectrum at pH 2.5 25 °C, showing the correlation peaks of the Hα proton at 5.37 ppm (green line). Their through correlation peaks (not indicated) led us to tentatively assign the Hα proton to Phe (B24). The correlation of the peaks at 6.6 and 7.09 ppm (blue lines) with the protons at 2.55 and 3.95 ppm probably refer to the signals of an asparagine residue.** ***Right*****: zoom in image of the selected area.**

At acidic pH, the peak at 6.28 ppm has a correlation with a peak at 7.49 ppm (Figure 2.21, in red). Given the combination of chemical shifts, the fact that no cross-peak was present in the ^{13}C-HSQC spectrum (Figure 2.22, red) and the exclusive presence at acidic pH, this likely corresponds to an arginine side-chain NH proton. From the peak at 7.49 ppm and its through-bond correlations we found that these peaks are related to protons at 2.96 (δ-CH_2), 1.66, 1.41 (γ-CH_2), 1.76, 1.86 (β-CH_2), 4.3 (Hα), 8.26 ppm (backbone H^N) (Figure 2.21, red lines). Thus the signal at 6.28 ppm is likely to refer to the side chain NH_2 protons of Arg (B22).

The peak at 6.60 ppm exhibits a cross-peak that correlates with the peak at 7.09 ppm in the 2D-TOCSY spectrum (Figure 2.21, blue). The absence of correlation peaks in the ^{13}C-HSQC spectrum (Figure 2.22, blue) indicates that both peaks also refer to side-chain NH protons, while both the 6.60 ppm and 7.09 ppm signals further correlate with the protons at 3.95 and 2.55 ppm in the 2D-NOESY spectrum (Figure 2.20, blue). Therefore they are likely to be signals of an asparagine residue.

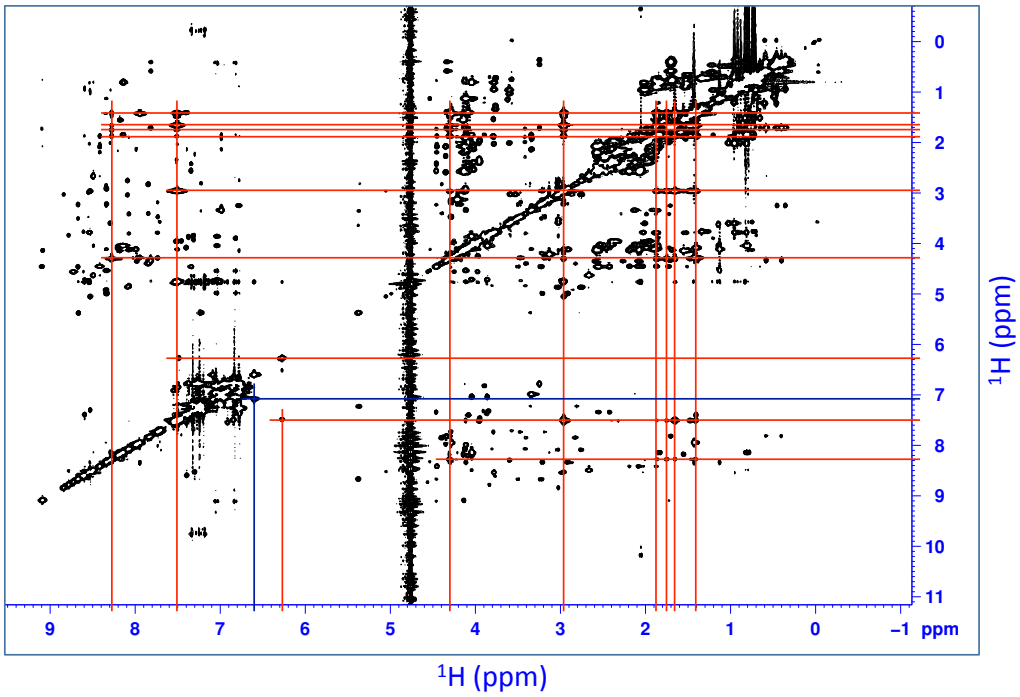


Figure 2.21: Full 2D-TOCSY spectrum at pH 2.5 25 °C, showing the correlation of the peak at 6.28 ppm with the peak at 7.49 ppm and the through-bond correlations of the peak at 7.49 ppm with the peaks at 1.41, 1.66 (γ-CH_2), 1.76, 1.86 (β-CH_2), 2.96 (δ-CH_2), 4.3 (Hα) and 8.26 ppm (backbone H^N) in red, probably matching an arginine residue. The correlation of the peak at 6.6 ppm with the peak at 7.09 ppm is shown in blue.

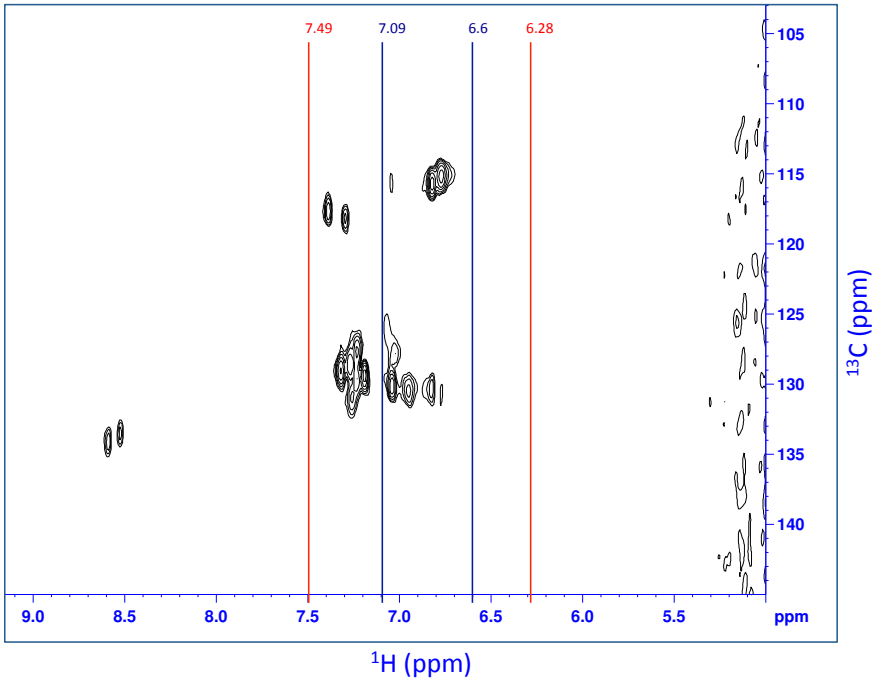


Figure 2.22: 2D-HSQC (^{13}C and ^{1}H) spectrum at pH 2.5 25 °C, showing the absence of cross-peaks for the peaks at 6.28, 7.49 (red) and 6.6, 7.09 (blue) therefore indicating that these peaks refer to side chain NH protons.

2.8. MOLECULAR MODELLING

The molecular modelling simulations were carried out by Dr. Shirin Jamshidi. In the molecular dynamics simulations, the monomer, dimer and hexamer were separately investigated for their dynamical behaviour as a function of increasing temperature. We used higher temperatures than in the laboratory to compensate for the shorter time scale of the simulations.

Monomer: The overlaid structures at 20, 50, 65, 75 and 110 °C (Figure 2.23i) illustrate increasing flexibility and progressive displacement of the B chain C-terminus,[126] and increasingly flexible N-termini for both A and B chains, while the α-helix content seems to be significantly reduced and the turn content increased at the end of the simulation (Table 2.7).

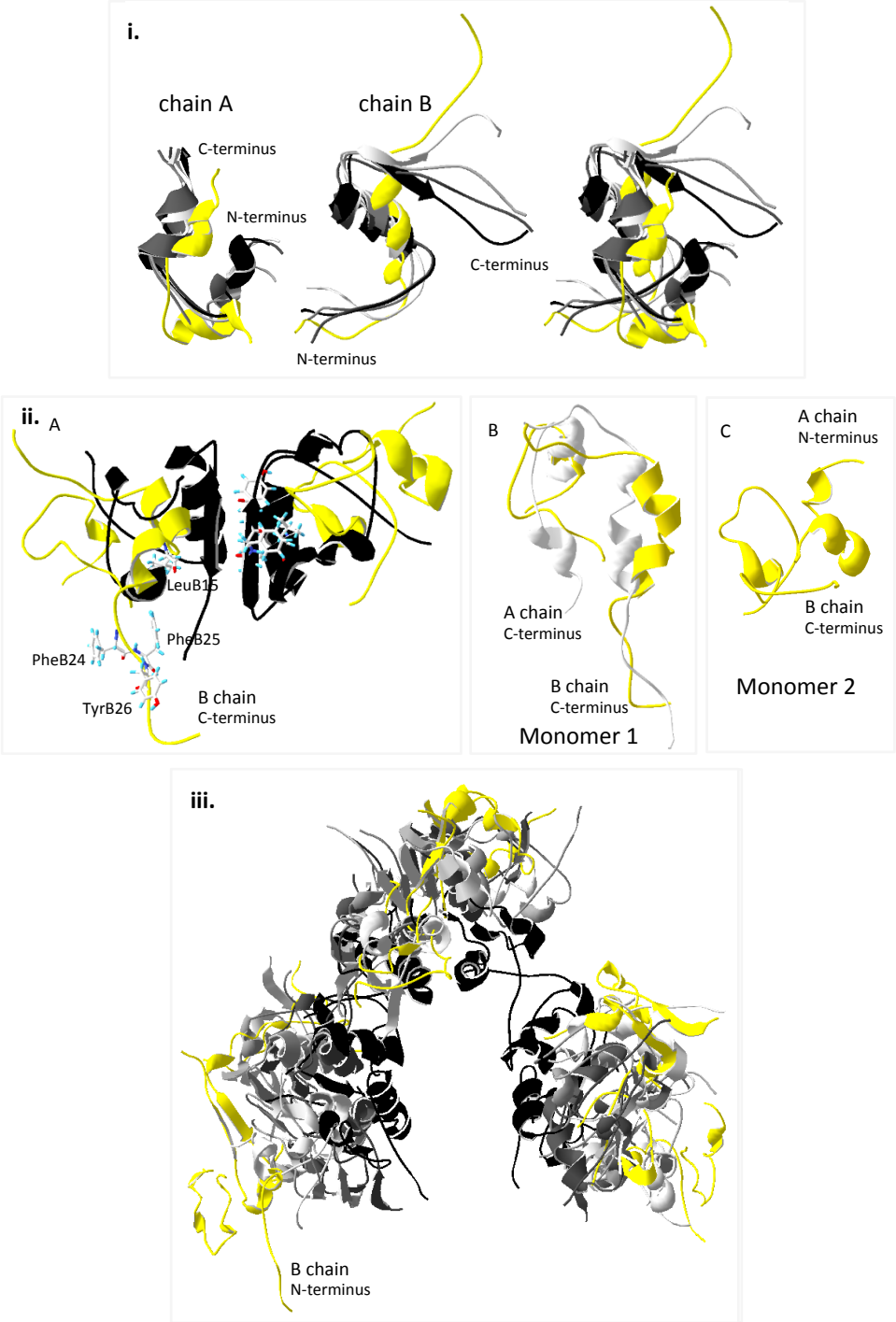


Figure 2.23: Overlaid structures from the SA-MD simulations at increasing temperature. From black to light grey with decreasing brightness, the frames are at 20, 50, 65 and 75 °C. The frame at 110 °C is shown in yellow. i. The *monomer* is depicted. Left: A chain, centre: B chain, right: A and B chains. ii. *Dimer*: A. The frames are at 20 °C (black) and 110 °C (yellow). Leu (B15) and the aromatic residues at the C-terminus of B chain are shown in both structures, indicating the increased distance at 110 °C. B. Comparison between monomer 1 from the dimer simulation at 110 °C (yellow) and the monomer simulation at 110 °C (white). The similarity of the two unfolded structures is apparent. C. Monomer 2 at 110 °C ended up in a differently unfolded conformation. iii. *Hexamer*: The expansion of the structure and the tendency towards dissociation with increasing temperature is shown.

Dimer: Figure 2.23ii A shows the overlaid structures at 20 and 110 °C. The dimer appears to dissociate at 110 °C, resulting in two differently misfolded monomers. The overlay of the 110 °C monomer structure with monomer 1 from the dimer simulation (Figure 2.23ii B) shows that the final structures are similar. The extended B chain C-terminus (which turns the Leu (B15) methyl group away from the aromatic residues Phe (B24), Phe (B25) and Tyr (B26)) leads us to the conclusion that this structure represents the completely unfolded conformation, while monomer 2 appears in a differently unfolded conformation (Figure 2.23ii C) probably referring to the partially unfolded intermediate. Figure 2.24 shows monomer 2 at 20, 50, 65, 75 and 110 °C. For both chains (A on the left and B on the right), which are depicted separately for clarity, the flexibility of the N-termini throughout the simulation is apparent. The displacement of the two chains at 110 °C (in yellow) is indicative of the high tendency for dissociation of the dimer to the two misfolded monomers.

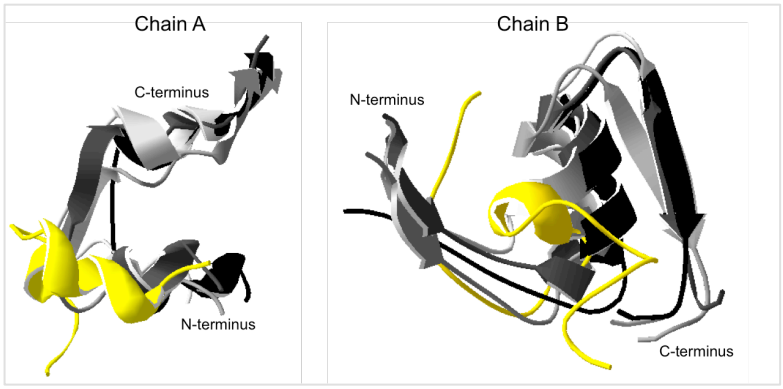


Figure 2.24: Overlaid structures from the dimer simulation (monomer 2) as a function of temperature for chain A (*Left*) and chain B (*Right*). From black to light gray with decreasing brightness, the frames are at 20, 50, 65 and 75 °C. The frame at 110 °C is shown in yellow.

Hexamer: Figure 2.23iii depicts the overlaid structures at 20, 50, 65, 75 and 110 °C. The structure gradually expands as the temperature increases, suggesting a tendency towards dissociation. At 110 °C the interactions between the antiparallel β-strands of the B chain C-terminal portions are weakened and the B chain N-terminus is completely extended. Each one of the three dimers is separately depicted in Figure 2.25, suggesting that each dimer can potentially follow a different unfolding pattern. Both A (left) and B (right) chains of dimer 1 appear to be more stable and compact in contrast to the increased flexibility of dimers 2 and 3, which gradually diverge leading to more expanded structures and to a complete extension of the N-termini of the B chains in the case of dimer 3. However, dimer 3 appears to preserve its antiparallel β-strand character in the same way as dimer 1, while dimer 2 appears prone to further

dissociation to monomers as the antiparallel β-strand character has disappeared. At the end of this simulation, a significant amount of helix content is still maintained (Table 2.7) which indicates that despite the evident onset of dissociation into monomers, the latter do not completely unfold on the short timescale of the simulations.

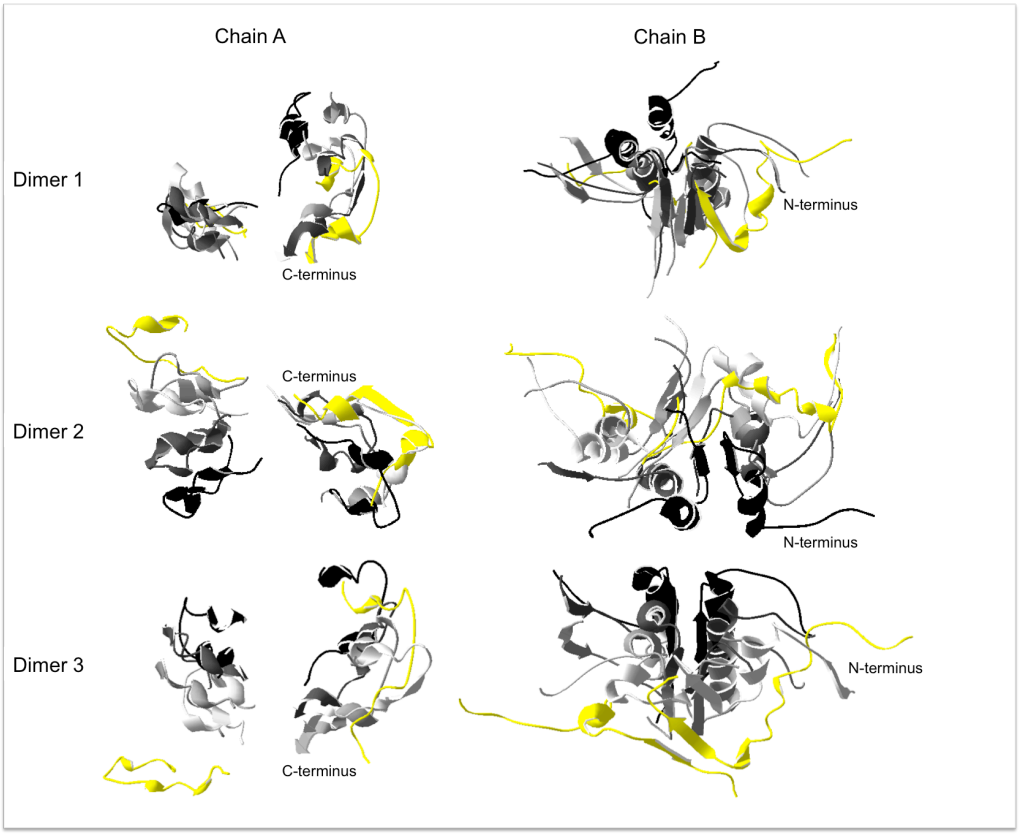


Figure 2.25: Structures from the hexamer simulation as a function of increasing temperature. Chains A (left) and B (right) are depicted separately for each of the three dimers. From black to light grey with decreasing brightness, the frames are at 20, 50, 65 and 75 °C. The frame at 110 °C is shown in yellow. The three dimers present differences in their unfolding patterns, therefore contributing to the increased stability of the hexamer, which needs significantly longer time and higher temperature in order to dissociate and unfold compared to the dimer.

The results from the secondary structure analysis for the monomer, dimer and hexamer simulations at 20, 50, 65, 75 and 110 °C, using the DSSP algorithm, are presented in Table 2.7. As mentioned above, the monomer and the dimer show a decrease in the α-helix content, while the hexamer seems to maintain a significant amount of its helical content at the end of the simulation. In all three cases the β-sheet content is hardly changing, which is contradicting the results from the rest of the techniques. This inaccuracy perhaps results from the fact that the simulations depict only one molecule in each case, while β-sheet formation is an intermolecular event caused by the aggregation of several molecules.

Table 2.7: Secondary structure analysis results from the simulations of insulin monomer, dimer and hexamer using the DSSP algorithm of the AMBER package. The results were averaged over three snapshots at each one of the shown temperatures. The abbreviations stand for: (h) helix, (s) strand, (t) turn, (o) other.

		20 °C	50 °C	65 °C	75 °C	110 °C
Monomer	(h)	0.45	0.46	0.36	0.41	0.20
	(s)	0.00	0.05	0.08	0.03	0.03
	(t)	0.13	0.18	0.20	0.17	0.37
	(o)	0.42	0.31	0.36	0.39	0.40
Dimer	(h)	0.40	0.38	0.38	0.34	0.27
	(s)	0.05	0.07	0.04	0.09	0.05
	(t)	0.17	0.18	0.21	0.24	0.31
	(o)	0.38	0.37	0.37	0.33	0.37
Hexamer	(h)	0.42	0.43	0.37	0.42	0.36
	(s)	0.02	0.03	0.07	0.03	0.05
	(t)	0.16	0.17	0.21	0.18	0.22
	(o)	0.41	0.38	0.35	0.38	0.37

2.9. CONCLUSION

The main conclusion from this work is that empirical comparisons of data collected with the complementary techniques of DLS, CD, FT-IR, NMR and molecular modelling coupled with innovative use of temperature dependence can be used to identify the occurrences of differences in structure in solution even when the samples are heterogeneous. This is therefore the beginning of the development of a batch-to-batch comparison methodology for comparing the similarity of a new formulation to the original product as part of the Quality Control procedure, which will ensure the safety and efficacy of a biopharmaceutical product.[127] DLS gives a simple indication of particle size which helps understanding of line-width variations in NMR spectra. Chemical shifts in different regions of the NMR spectrum give evidence for structural changes but more detailed analysis of methyl protons, histidines, or other aromatic protons (particularly when coupled with temperature dependence) enabled us to monitor the presence or loss of intermolecular interactions as well as stabilities of samples. CD gave a picture of structural differences and the relative stabilities of different samples which could be plotted as percentage changes of different secondary structure motifs as a function of temperature. Along with CD spectroscopy, FT-IR can look into the secondary structure of globular proteins, showing a promising use in combination with our retrained SSNN software, and probably is the first choice technique in studying aggregation and fibrillation phenomena. Although the simulations were only run for a very short time (ns), they provided a useful complement to the experiments when considered over a greater temperature range than the experiments, letting us visualise which parts of the molecules were likely to unfold first.

The significance of temperature as a variable is that we were able to observe quite subtle differences in structure and/or stability by seeing the different temperatures at which the structural motifs were lost and β-strand structure appeared. There was no evidence of increased β-strand structure in the molecular dynamics simulations at higher temperatures, so we conclude that the β-strand formation is an intermolecular event that reflects the early stages of the process that later leads to fibril formation and is favoured at higher concentrations such as those used for DLS, FT-IR and NMR. Thus combining temperature dependent DLS, CD (with structure analysis), and molecular modelling where possible should be explored for probing early aggregate formation. Although most of our data confirmed and consolidated existing literature

data for insulin, we are not aware of any combined use of these techniques in conjunction with temperature variation to probe the early stages of fibre formation.

References

1. Nobel Media AB, "The Nobel Prize in Physiology or Medicine 1923", http://www.nobelprize.org/nobel_prizes/medicine/laureates/1923/, (accessed 16 Jan, 2017).
2. Nobel Media AB, "The Nobel Prize in Chemistry 1958", https://www.nobelprize.org/nobel_prizes/chemistry/laureates/1958/, (accessed 16 Jan, 2017).
3. J.-H. Wang, *Trends in Biochemical Sciences*, **23**, 497-500.
4. L. K. Altman, A new insulin given approval for use in U.S., http://www.nytimes.com/1982/10/30/us/a-new-insulin-given-approval-for-use-in-us.html, (accessed 4th October 2016).
5. J. H. DeVries, S. C. L. Gough, J. Kiljanski and L. Heinemann, *Diabetes, Obesity & Metabolism*, 2015, **17**, 445-451.
6. M. Eisenstein, *Nature Biotechnology*, 2011, **29**, 782-785.
7. C. Sorli, *The American Journal of Medicine*, **127**, S39-S48.
8. European Medicines Agency, Marvel LifeSciences Ltd withdraws its marketing authorisation applications for Solumarv, Isomarv and Combimarv (human insulin), http://www.ema.europa.eu/docs/en_GB/document_library/Press_release/2012/11/WC500135156.pdf, (accessed 26 Oct, 2015).
9. Y. Pocker and S. B. Biswas, *Biochemistry*, 1980, **19**, 5043-5049.
10. S. N. Timasheff, M. J. Ettinger and M. G. D. Strycharz, *Biochemistry*, 1971, **10**, 824-831.
11. J. Goldman and F. H. Carpenter, *Biochemistry*, 1974, **13**, 4566-4574.
12. J. H. Bradbury, V. Ramesh and G. Dodson, *Journal of Molecular Biology*, 1981, **150**, 609-613.
13. J. H. Bradbury and V. Ramesh, *Biochemical Journal*, 1985, **229**, 731-737.
14. R. Palmieri, R. W. K. Lee and M. F. Dunn, *Biochemistry*, 1988, **27**, 3387-3397.
15. M. Roy, R. W. Lee, J. Brange and M. F. Dunn, *Journal of Biological Chemistry*, 1990, **265**, 5448-5452.
16. M. Roy, R. W. K. Lee, N. C. Kaarsholm, H. Thøgersen, J. Brange and M. F. Dunn, *Biochimica and Biophysica Acta, Molecular Cell Research*, 1990, **1053**, 63-73.
17. W. Kadima, M. Roy, R. W. Lee, N. C. Kaarsholm and M. F. Dunn, *Journal of Biological Chemistry*, 1992, **267**, 8963-8970.
18. W. Bocian, J. Sitkowski, E. Bednarek, A. Tarnowska, R. Kawęcki and L. Kozerski, *Journal of Biomolecular NMR*, 2008, **40**, 55-64.
19. J. Kraineva, V. Smirnovas and R. Winter, *Langmuir*, 2007, **23**, 7118-7126.
20. F. G. Backlund, J. Wigenius, F. Westerlund, O. Inganas and N. Solin, *Journal of Materials Chemistry C*, 2014, **2**, 7811-7822.

21. W. Dzwolak, R. Ravindra and R. Winter, *Physical Chemistry Chemical Physics*, 2004, **6**, 1938-1943.
22. W. Dzwolak, R. Ravindra, J. Lendermann and R. Winter, *Biochemistry*, 2003, **42**, 11347-11355.
23. T. Sneideris, D. Darguzis, A. Botyriute, M. Grigaliunas, R. Winter and V. Smirnovas, *PLOS ONE*, 2015, **10**, e0136602.
24. S. Ma, X. Cao, M. Mak, A. Sadik, C. Walkner, T. B. Freedman, I. K. Lednev, R. K. Dukor and L. A. Nafie, *Journal of the American Chemical Society*, 2007, **129**, 12364-12365.
25. W. Kadima, L. Øgendal, R. Bauer, N. Kaarsholm, K. Brodersen, J. F. Hansen and P. Porting, *Biopolymers*, 1993, **33**, 1643-1657.
26. S. Hvidt, *Biophysical Chemistry*, 1991, **39**, 205-213.
27. A. K. Attri, C. Fernández and A. P. Minton, *Biophysical Chemistry*, 2010, **148**, 28-33.
28. H. B. Bohidar, *Biopolymers*, 1998, **45**, 1-8.
29. Y. Xu, Y. Yan, D. Seeman, L. Sun and P. L. Dubin, *Langmuir*, 2012, **28**, 579-586.
30. G. D. Smith, W. A. Pangborn and R. H. Blessing, *Acta Crystallographica Section D: Biological Crystallography*, 2005, **61**, 1476-1482.
31. C. G. Frankær, M. V. Knudsen, K. Norén, E. Nazarenko, K. Ståhl and P. Harris, *Acta Crystallographica Section D*, 2012, **68**, 1259-1271.
32. I. Margiolaki, A. E. Giannopoulou, J. P. Wright, L. Knight, M. Norrman, G. Schluckebier, A. N. Fitch and R. B. Von Dreele, *Acta Crystallographica Section D*, 2013, **69**, 978-990.
33. O. Gursky, J. Badger, Y. Li and D. L. Caspar, *Biophysical Journal*, **63**, 1210-1220.
34. E. J. Nettleton, P. Tito, M. Sunde, M. Bouchard, C. M. Dobson and C. V. Robinson, *Biophysical Journal*, 2000, **79**, 1053-1065.
35. D. Fabris and C. Fenselau, *Analytical Chemistry*, 1999, **71**, 384-387.
36. E. N. Baker, T. L. Blundell, J. F. Cutfield, S. M. Cutfield, E. J. Dodson, G. G. Dodson, D. M. Hodgkin, R. E. Hubbard, N. W. Isaacs, C. D. Reynolds and et al., *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, 1988, **319**, 369-456.
37. F. Sanger, *Biochemical Journal*, 1949, **44**, 126-128.
38. F. Sanger and H. Tuppy, *Biochemical Journal*, 1951, **49**, 463-481.
39. F. Sanger and H. Tuppy, *Biochemical Journal*, 1951, **49**, 481-490.
40. F. Sanger and E. O. P. Thompson, *Biochemical Journal*, 1953, **53**, 353-366.
41. F. Sanger and E. O. P. Thompson, *Biochemical Journal*, 1953, **53**, 366-374.
42. H. R. Morris and P. Pucci, *Biochemical and Biophysical Research Communications*, 1985, **126**, 1122-1128.
43. M. L. Brader, N. C. Kaarsholm, R. W. K. Lee and M. F. Dunn, *Biochemistry*, 1991, **30**, 6636-6645.
44. M. L. Dieken, M. Federwisch, P. De Meyts and A. Wollmer, *Insulin & Related Proteins — Structure to Function and Pharmacology*, Springer Netherlands, 2002.

45. B. Skelbaek-Pedersen, J. Brange, L. Langkjaer, U. Damgaard, H. Ege, S. Havelund, L. G. Heding, K. H. Joergensen, J. Lykkeberg and J. Markussen, *Galenics of Insulin: The Physico-chemical and Pharmaceutical Aspects of Insulin and Insulin Preparations*, Springer Berlin Heidelberg, 2012.
46. A. H. Pekar and B. H. Frank, *Biochemistry*, 1972, **11**, 4013-4016.
47. P. D. Jeffrey and J. H. Coates, *Biochemistry*, 1966, **5**, 489-498.
48. S. Ludvigsen, M. Roy, H. Thoegersen and N. C. Kaarsholm, *Biochemistry*, 1994, **33**, 7998-8006.
49. M. J. Adams, T. L. Blundell, E. J. Dodson, G. G. Dodson, M. Vijayan, E. N. Baker, M. M. Harding, D. C. Hodgkin, B. Rimmer and S. Sheat, *Nature*, 1969, **224**, 491-495.
50. T. Blundell, G. Dodson, D. Hodgkin and D. Mercola, in *Advances in Protein Chemistry*, eds. J. T. E. C.B. Anfinsen and M. R. Frederic, Academic Press, 1972, vol. Volume 26, pp. 279-402.
51. J. Brange, *Stability of Insulin*, Springer, 1994.
52. G. Dodson and D. Steiner, *Current Opinion in Structural Biology*, 1998, **8**, 189-194.
53. J. Brange, U. Ribel, J. F. Hansen, G. Dodson, M. T. Hansen, S. Havelund, S. G. Melberg, F. Norris, K. Norris, L. Snel, A. R. Sorensen and H. O. Voigt, *Nature*, 1988, **333**, 679-682.
54. F. Y. H. Lin, D. Y. Kwok, Z. Policova, W. Zingg and A. W. Neumann, *Colloids and Surfaces, B*, 1995, **3**, 281-286.
55. J. Gavrilova, V. Tougu and P. Palumaa, *Metallomics*, 2014, **6**, 1296-1300.
56. S. G. Melberg and W. C. Johnson, *Proteins: Structure, Function and Bioinformatics*, 1990, **8**, 280-286.
57. Z.-P. Yao, Z.-H. Zeng, H.-M. Li, Y. Zhang, Y.-M. Feng and D.-C. Wang, *Acta Crystallographica Section D*, 1999, **55**, 1524-1532.
58. L. Schrödinger, *The PyMOL Molecular Graphics System, Version 1.74*, 2015.
59. R. E. Ratner, T. M. Phillips and M. Steiner, *Diabetes*, 1990, **39**, 728-733.
60. L. N. Garriques, S. Frokjaer, J. F. Carpenter and J. Brange, *Journal of Pharmaceutical Sciences*, 2002, **91**, 2473-2480.
61. R. J. Woods, J. Alarcón, E. McVey and R. J. Pettis, *Journal of Diabetes Science and Technology*, 2012, **6**, 265-276.
62. D. R. Owens, *Diabetes Technology & Therapeutics*, 2011, **13**, S-5-S-14.
63. W. Oakley, D. Hill and N. Oakley, *Diabetes*, 1966, **15**, 219-222.
64. World Health Organisation, WHO International Standard Insulin Human, https://www.nibsc.org/documents/ifu/83-500.pdf, (accessed 17 Jan, 2017).
65. European Medicines Agency, Results for active substance insulin, http://www.ema.europa.eu/ema/index.jsp?curl=pages%2Fmedicines%2Flanding%2Fepar_search.jsp&mid=WC0b01ac058001d124&searchTab=searchByKey&alreadyLoaded=true&isNewQuery=true&status=Authorised&status=Withdrawn&status=Suspended&status=Refused&keyword=insulin&keywordSearch=Submit&searchType=inn&taxonomyPath=&treeNumber=&searchGenericType=generics, (accessed 17 Jan, 2017).

66. E. Blair, Insulin A to Z: A Guide on Different Types of Insulin, http://www.joslin.org/info/insulin_a_to_z_a_guide_on_different_types_of_insulin.html, (accessed 17 Jan, 2017).
67. J. Sandow, W. Landgraf, R. Becker and G. Seipke, *European Journal of Endocrinology*, 2015, **11**, 10-16.
68. Diabetes.co.uk, Insulin Types and Information, http://www.diabetes.co.uk/insulin/insulin-types.html, (accessed 17 Jan, 2017).
69. L. Weil, T. S. Seibles and T. T. Herskovits, *Archives of Biochemistry and Biophysics*, 1965, **111**, 308-320.
70. V. Hall, A. Nash and A. Rodger, *Analytical Methods*, 2014, **6**, 6721-6726.
71. N. Sreerama and R. W. Woody, *Analytical Biochemistry*, 1993, **209**, 32-44.
72. T. L. Hwang and A. J. Shaka, *Journal of Magnetic Resonance, Series A*, 1995, **112**, 275-279.
73. D. Wishart, C. Bigam, J. Yao, F. Abildgaard, H. J. Dyson, E. Oldfield, J. Markley and B. Sykes, *Journal of Biomolecular NMR*, 1995, **6**, 135-140.
74. A. Bax and D. G. Davis, *Journal of Magnetic Resonance (1969)*, 1985, **65**, 355-360.
75. A. G. Palmer, J. Cavanagh, P. E. Wright and M. Rance, *Journal of Magnetic Resonance (1969)*, 1991, **93**, 151-170.
76. L. Kay, P. Keifer and T. Saarinen, *Journal of the American Chemical Society*, 1992, **114**, 10663-10665.
77. J. Schleucher, M. Schwendinger, M. Sattler, P. Schmidt, O. Schedletzky, S. J. Glaser, O. W. Sørensen and C. Griesinger, *Journal of Biomolecular NMR*, 1994, **4**, 301-306.
78. E. L. Gilroy, M. R. Hicks, D. J. Smith and A. Rodger, *Analyst*, 2011, **136**, 4159-4163.
79. T. Mokhtari, Doctor of Philosophy, Kansas State University, 2007.
80. V. Tsui and D. A. Case, *Biopolymers*, 2000, **56**, 275-291.
81. T. Darden, D. York and L. Pedersen, *Journal of Chemical Physics*, 1993, **98**, 10089-10092.
82. J.-P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, *Journal of Computational Physics*, 1977, **23**, 327-341.
83. D. A. Case, T. E. Cheatham, 3rd, T. Darden, H. Gohlke, R. Luo, K. M. Merz, Jr., A. Onufriev, C. Simmerling, B. Wang and R. J. Woods, *Journal of Computational Chemistry*, 2005, **26**, 1668-1688.
84. W. Kabsch and C. Sander, *Biopolymers*, 1983, **22**, 2577-2637.
85. A. Micsonai, F. Wien, L. Kernya, Y.-H. Lee, Y. Goto, M. Réfrégiers and J. Kardos, *Proceedings of the National Academy of Sciences*, 2015, **112**, E3095-E3103.
86. N. Sreerama, S. Y. Venyaminov and R. W. Woody, *Protein Science : A Publication of the Protein Society*, 1999, **8**, 370-380.
87. N. Guex and M. C. Peitsch, *Electrophoresis*, 1997, **18**, 2714-2723.
88. A. K. Attri, C. Fernández and A. P. Minton, *Biophysical Chemistry*, 2010, **148**, 23-27.
89. L. Hovgaard, H. Jacobs, N. A. Mazer and S. W. Kim, *International Journal of Pharmaceutics*, 1996, **132**, 107-113.

90. B. Nordén, A. Rodger, T. Dafforn, *Linear Dichroism and Circular Dichroism: A Textbook on Polarized-light Spectroscopy*, Royal Society of Chemistry, 2010.
91. J. W. S. Morris, D. A. Mercola and E. R. Arquilla, *Biochimica et Biophysica Acta, Protein Structure*, 1968, **160**, 145-150.
92. L. Whitmore and B. A. Wallace, *Nucleic Acids Research*, 2004, **32**, W668-W673.
93. Y. Zhang, J. L. Whittingham, J. P. Turkenburg, E. J. Dodson, J. Brange and G. G. Dodson, *Acta Crystallographica Section D*, 2002, **58**, 186-187.
94. V. Hall, A. Nash, E. Hines and A. Rodger, *Journal of Computational Chemistry*, 2013, **34**, 2774-2786.
95. V. N. Uversky and A. L. Fink, *Biochimica et Biophysica Acta*, 2004, **1698**, 131-153.
96. M. R. H. Krebs, E. H. C. Bromley, S. S. Rogers and A. M. Donald, *Biophysical Journal*, 2005, **88**, 2013-2021.
97. A. Arora, C. Ha and C. B. Park, *Protein Science*, 2004, **13**, 2429-2436.
98. Q.-x. Hua and M. A. Weiss, *Journal of Biological Chemistry*, 2004, **279**, 21449-21460.
99. S. M. Kelly, T. J. Jess and N. C. Price, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 2005, **1751**, 119-139.
100. A. J. Miles and B. A. Wallace, *Chemical Society Reviews*, 2006, **35**, 39-51.
101. O. Bozkurt, S. Haman Bayari, M. Severcan, C. Krafft, J. Popp and F. Severcan, *Journal of Biomedical Optics*, 2012, **17**, 0760231-0760238.
102. J. A. Hering, P. R. Innocent and P. I. Haris, *Proteomics*, 2004, **4**, 2310-2319.
103. M. Severcan, F. Severcan and P. I. Haris, *Journal of Molecular Structure*, 2001, **565–566**, 383-387.
104. Y. Wang, R. I. Boysen, B. R. Wood, M. Kansiz, D. McNaughton and M. T. W. Hearn, *Biopolymers*, 2008, **89**, 895-905.
105. J. C. Gorga, A. Dong, M. C. Manning, R. W. Woody, W. S. Caughey and J. L. Strominger, *Proceedings of the National Academy of Sciences*, 1989, **86**, 2321-2325.
106. A. Dong, P. Huang and W. S. Caughey, *Biochemistry*, 1990, **29**, 3303-3308.
107. J. Kong and S. Yu, *Acta Biochimica et Biophysica Sinica*, 2007, **39**, 549-559.
108. A. Bellomaria, R. Nepravishta, M. Marchetti and M. Paci, *Food Chemistry*, 2016, **194**, 733-739.
109. M. B. Taraban, H. C. Truong, Y. Feng, E. V. Jouravleva, M. A. Anisimov and Y. B. Yu, *Journal of Pharmaceutical Sciences*, 2015, **104**, 4132-4141.
110. M. F. Lin and C. K. Larive, *Analytical Biochemistry*, 1995, **229**, 214-220.
111. R. Page, W. Peti, I. A. Wilson, R. C. Stevens and K. Wüthrich, *Proceedings of the National Academy of Sciences U. S. A.*, 2005, **102**, 1901-1905.
112. V. Ramesh and J. H. Bradbury, *Archives of Biochemistry and Biophysics*, 1987, **258**, 112-122.
113. K. A. Higgins, D. J. Craik and J. G. Hall, *Biochemistry International*, 1990, **22**, 627-637.
114. A. D. Kline and R. M. Justice, *Biochemistry*, 1990, **29**, 2906-2913.

115. T. Blundell, G. Dodson, D. Hodgkin and D. Mercola, in *Advances in Protein Chemistry*, eds. J. T. E. C.B. Anfinsen and M. R. Frederic, Academic Press, 1972, vol. Volume 26, pp. 279-402.
116. N. J. Baxter, L. L. P. Hosszu, J. P. Waltho and M. P. Williamson, *Journal of Molecular Biology*, 1998, **284**, 1625-1639.
117. K. Huang, B. Xu, S. Q. Hu, Y. C. Chu, Q. X. Hua, Y. Qu, B. Li, S. Wang, R. Y. Wang, S. H. Nakagawa, A. M. Theede, J. Whittaker, P. De Meyts, P. G. Katsoyannis and M. A. Weiss, *Journal of Molecular Biology*, 2004, **341**, 529-550.
118. K. Huang, S. J. Chan, Q. X. Hua, Y. C. Chu, R. Y. Wang, B. Klaproth, W. Jia, J. Whittaker, P. De Meyts, S. H. Nakagawa, D. F. Steiner, P. G. Katsoyannis and M. A. Weiss, *Journal of Biological Chemistry*, 2007, **282**, 35337-35349.
119. J. G. Menting, Y. Yang, S. J. Chan, N. B. Phillips, B. J. Smith, J. Whittaker, N. P. Wickramasinghe, L. J. Whittaker, V. Pandyarajan, Z. Wan, S. P. Yadav, J. M. Carroll, N. Strokes, C. T. Roberts, F. Ismail-Beigi, W. Milewski, D. F. Steiner, V. S. Chauhan, C. W. Ward, M. A. Weiss and M. C. Lawrence, *Proceedings of the National Academy of Sciences*, 2014, **111**, E3395-E3404.
120. S. M. Kristensen, A. M. M. Jørgensen, J. J. Led, P. Balschmidt and F. B. Hansen, *Journal of Molecular Biology*, 1991, **218**, 221-231.
121. S. P. Mielke and V. V. Krishnan, *Progress in Nuclear Magnetic Resonance Spectroscopy*, 2009, **54**, 141-165.
122. D. S. Wishart, B. D. Sykes and F. M. Richards, *FEBS Lett.*, 1991, **293**, 72-80.
123. R. Tycko and R. B. Wickner, *Accounts of Chemical Research*, 2013, **46**, 1487-1496.
124. M. Roy, M. L. Brader, R. W. Lee, N. C. Kaarsholm, J. F. Hansen and M. F. Dunn, *J. Biol. Chem.*, 1989, **264**, 19081-19085.
125. P. S. Brzovic, W. E. Choi, D. Borchardt, N. C. Kaarsholm and M. F. Dunn, *Biochemistry*, 1994, **33**, 13057-13069.
126. V. Zoete, M. Meuwly and M. Karplus, *Journal of Molecular Biology*, 2004, **342**, 913-929.
127. European Medicines Agency, Note for guidance on specifications: Test procedures and acceptance criteria for biotechnological/biological products (CPMP/ICH/365/96) http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002824.pdf, (accessed 28 Oct, 2015).

Abstract

Monoclonal antibodies and antibody-drug conjugates are state-of-the-art biotherapeutics against various types of cancer. However, the complexity of their structure induces high lot-to-lot heterogeneity depending on the manufacturing conditions, and can affect the stability, function and immunogenicity of the final product. EMA guidelines require that homogeneity between samples should be demonstrated in cases where biosimilar-to-innovator and batch-to-batch comparison is pursued. In this work, two native samples of trastuzumab (mAbs) were compared with two conjugated samples (ADCs), while at the same time comparison of the used conjugation methods was carried out. Mass spectrometry (MS), circular dichroism (CD) and dynamic light scattering (DLS) in combination, and coupled with temperature dependence, were able to cover the primary structure, post-translational modifications (PTMs), relative drug-conjugation, secondary structure, stability, purity and aggregation of the samples. Overall, all four samples appeared stable at least up to 54 °C; ADCs presented reduced stability compared to the mAbs because of their more flexible hinge region, while the new patented conjugation method appeared more controlled and yielded products of higher purity in comparison to the traditional technique.

Chapter 3
Monoclonal Antibodies (mAbs) and Antibody-Drug Conjugates (ADCs)

3.1. INTRODUCTION

Antibodies are glycoproteins secreted by B cells as a response to substances that are recognised as foreign by the organism (antigens). The antibody-antigen interaction is characterised by high specificity (lock and key concept), which makes antibodies ideal molecules for targeted therapies. In the early 1900s, Paul Ehrlich was the first to express the idea of molecules that would be able to deliver the drug directly to the disease site,[1] receiving a Nobel Prize in Physiology or Medicine in 1908. However, the "magic bullets" concept did not see much progress due to difficulties in isolating and purifying antibodies from animal or human serum. Milstein and Köhler were the first to successfully produce a monoclonal* antibody (mAb) in 1975,[2] while Muromonab-CD3 (OKT3) was the first antibody to receive FDA approval in 1986 as immunosuppressive agent for kidney transplantation.[3] Murine mAbs, however, had the disadvantage of causing allergic reactions followed by triggering of anti-drug antibodies (ADAs). With the advent of recombinant technology, the production of humanlike antibodies became possible, initially as chimeric[4] (65% human) and later as humanised[5] (95% human). Fully human mAbs followed with the development of *in vitro* phage-display technology[6-8] and the generation of mouse strains capable of expressing human variable domains.[9,10] Both humanised and fully human mAbs act similarly to endogenous immunoglobulin G (IgG), and have significantly reduced immunogenicity.

Nowadays, "magic bullets" refer to antibody-mediated targeted therapy using structures such as antibody-drug conjugates (ADCs), which combine the specificity of mAbs with enhanced potency, due to the attached cytotoxic agents. Unlike conventional anticancer therapeutics that can also attack healthy cells, usually resulting in increased toxicity, side effects and development of drug resistance, ADCs exploit the property of endocytosis and are designed to be inactive in the bloodstream. The antibodies recognise and bind to antigens that are specifically expressed by tumour

* Produced by a single cell line, in contrast with polyclonal antibodies, which are secreted by different B cell lines.

cells and are found on the target cells' surface. In that way the ADCs enter the cells (through endocytosis) and, after being processed, they release the attached drugs (toxins), which act either as antimitotic agents by inhibiting microtubule polymerisation[11-13] or by targeting DNA, cleaving the double-strand at specific sites,[14] and eventually lead to cell apoptosis. Gemtuzumab ozogamicin (Mylotarg) was the first ADC that received FDA approval in 2000 for acute myeloid leukemia,[15] however marketing authorisation was refused by EMA in 2008[16] and the drug was withdrawn from the U.S. market in 2010.[17] EMA has so far approved two ADCs: brentuximab vedotin (Adcetris) for Hodgkin lymphoma,[18,19] and trastuzumab emtansine (Kadcyla) for breast cancer.[20] Kadcyla is the recommended therapy to patients who have previously received the related mAb, trastuzumab—commercially available in the European market since 2000, under the name Herceptin.[21]

Such complex molecules, as biotherapeutics or more specifically monoclonal antibodies, require a series of tests in order to support comparability between a follow-on version (biosimilar) and the original innovator's product, as well as batch-to-batch similarity. That is because heterogeneity in biologics can occur in any stage of the production, from expression to purification, conjugation (when the molecule under question is an ADC) and storage. Changes in the conditions used in any of the above-mentioned steps might induce heterogeneity due to glycosylation variants, lysine-clipping variants[22,23] (both of which are referred to as post-translational modifications, PTMs), drug load variants (displaying different drug-to-antibody ratios), and modifications caused by degradation of the protein molecule such as deamidation, oxidation and aggregation. All mentioned modifications (PTMs, drug load variants and modifications resulting from degradation) can affect the stability, potency and immunogenicity of the final product. Full control of the manufacturing process is therefore of utmost importance, and accordingly significant is inspection of the structural similarity and the amount of impurities or degradation products among samples.

This chapter is a comparability study on the stability of four trastuzumab molecules, two of which are the native antibodies (mAbs) and the other two are conjugated to an antimitotic agent (ADCs). The two conjugates were produced using two different techniques: ADC (S) was produced with the traditional "in-solution" method,[24] while for ADC (LR) a patented "lock-release" method[25] was adopted. The sample called mAb (LR) went through the patented process including binding to a resin and release. The selected analytical techniques in this chapter follow the ICH

Guidelines for structural characterisation and confirmation of the molecule.[26] As in Chapter 2, the room temperature data were complemented with temperature dependence, which emphasises subtle differences that might be overlooked at room temperature.

Mass spectrometry (MS) was used as a tool to confirm the primary structure (amino acid sequence) of the antibody samples and also in order to look at post-translational modifications, such as carbohydrate structure and position, and degradation products (asparagine deamidation and methionine oxidation). Relative quantitation of the toxin molecules on each one of the conjugated samples was also applied. Dynamic light scattering (DLS) was chosen for its ability to show any impurities in the samples or formation of aggregates (especially at higher temperatures), and circular dichroism (CD) provides information on the secondary structure of the molecules and the loss of stability with the increase of temperature.

Commercial products of trastuzumab

As mentioned above, two products have been authorised—in both the EU and U.S. markets—based on the monoclonal antibody trastuzumab and are distributed by Roche. Herceptin (trastuzumab) is a recombinant DNA-derived humanised IgG1 monoclonal antibody produced by mammalian (Chinese hamster ovary) cell suspension culture and purified by affinity and ion exchange chromatography. Trastuzumab selectively binds to the human epidermal growth factor receptor 2 protein (HER2), inhibiting signalling and preventing proteolytic cleavage of its extracellular domain, which activates the receptor under normal conditions. As a result, proliferation of human tumour cells that overexpress HER2 is inhibited and Herceptin acts as an antineoplastic agent, indicated for breast and gastric cancer. The list of excipients, in the vial that is provided as a lyophilised powder* for reconstitution, includes L-histidine HCl, L-histidine, D-(+)-trehalose dihydrate and Tween 20 (similar to the samples provided for analysis in this study, see Materials). The mixture is reconstituted with 20 mL of bacteriostatic water for injection containing 1.1% benzyl alcohol as a preservative, yielding a ~21 mg/mL solution of trastuzumab at a pH of ~6. The solution is further diluted with 250 mL of 0.9% NaCl solution in order to be intravenously administered in personalised doses of 4 mg/kg body weight.[27]

* A concentrated solution for injection is also available, and in addition to the above-mentioned excipients it contains recombinant human hyaluronidase (rHuPH20) and L-methionine. It is administered subcutaneously after dilution with 0.9% or 0.45% NaCl solution.

Kadcyla (trastuzumab emtansine), is an antibody-drug conjugate of trastuzumab, covalently linked to a maytansine derivative (DM1), which is a microtubule inhibitor, via a stable thioether linker (MCC, 4-[N-maleimidomethyl]cyclohexane-1-carboxylate).[28] An average of 3.5 DM1 molecules are conjugated to each molecule of trastuzumab, and Kadcyla combines the action of both trastuzumab—as was described above—and the cytotoxic agent DM1. Selectivity of the antibody for the HER2-overexpressing tumour cells increases intracellular delivery of DM1 directly to the malignant cells. Upon internalisation, the ADC undergoes subsequent lysosomal degradation, resulting in release of DM1-containing cytotoxic catabolites and cell apoptosis. The list of excipients contains succinic acid, sodium hydroxide, sucrose and Tween 20, and the mixture is reconstituted with sterile water yielding a 20 mg/mL solution. The reconstituted solution is further diluted with 250 mL of 0.9% NaCl solution and is administered intravenously in personalised doses of 3.6 mg/kg body weight.[29] According to EMA, this medicine is subject to additional monitoring, and intensive studies are still in progress due to limited data on its long-term use.

3.1.1. Antibody structure

Antibodies are symmetric glycoproteins of the Immunoglobulin (Ig) superfamily, consisting of two identical heavy chains, of approximately 50 kDa each, and two equal light chains of 25 kDa each, held together by disulfide bonds (Figure 3.1). Each chain is divided in variable (V_L, V_H) and constant domains (C_L, C_H), with the variable domains determining the antigen-recognising part of the molecule and the constant domains determining the classes and subclasses of the antibody. While for the light chain there are only two types, defining an antibody as kappa or lambda, for the heavy chain there are five varieties that classify the antibody as Ig G, M, A, D and E. IgG constitutes 75% of the total immunoglobulins in the plasma of healthy individuals, therefore therapeutic monoclonal antibodies are typically of the IgG class. Depending on the heavy chain, IgG subclasses (or isotypes) can be defined as 1, 2, 3 and 4, with IgG1 being the most common (66%) and IgG2 following (23%). All four subclasses present >95% homology in the amino acid sequences of the constant domains of the heavy chains (C_H1, C_H2, C_H3), but differ in the hinge region both in amino acid composition and structure (length, and number of inter-heavy chain disulfide bridges). The variable regions for both the heavy and light chains consist of the complementary-determining regions (CDRs), which bind to the antigen, and the framework regions, which aid the CDR to contact the antigen. The variable and the attached to them

constant regions of both the light and heavy chains (V_L, C_L, V_H, C_H1) combine to form the fragment antigen binding domain (Fab), while the fragment crystallisable (Fc) domain responsible for triggering cell lysis is composed of the heavy chain C_H2 and C_H3 constant domains.

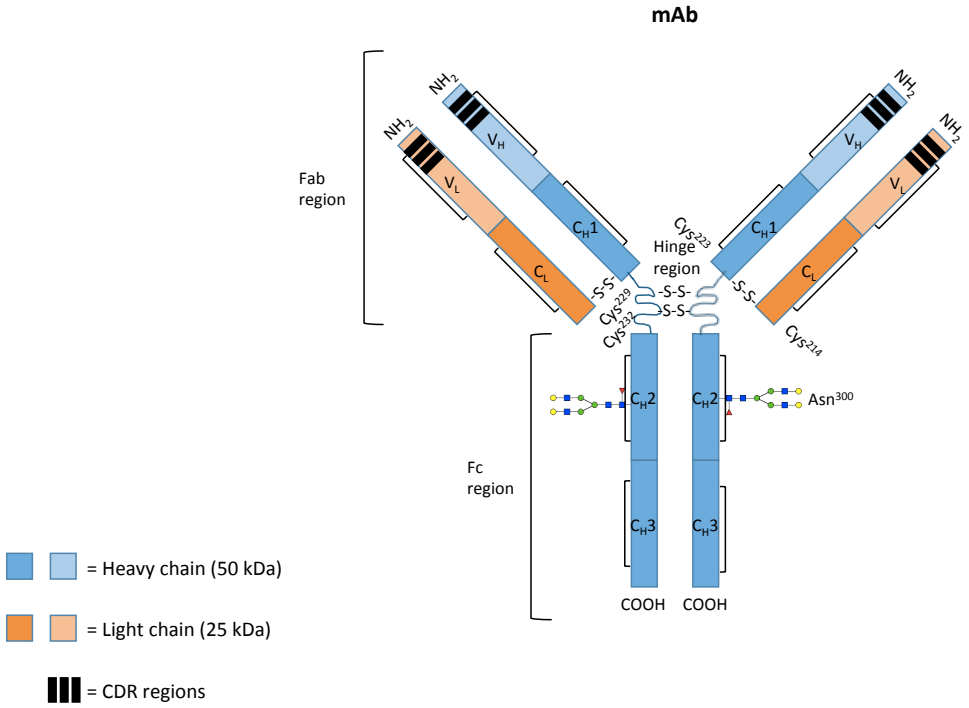


Figure 3.1: Representation of trastuzumab structure (or an IgG1 antibody), showing the variable domains (V_H, V_L) with the CDRs and the constant domains (C_H1–3, C_L) of the heavy (H) and light (L) chains. The intra-chain disulfide bonds are indicated with black lines for each domain, while the cysteine (Cys) residues forming inter-chain disulfide bonds are labelled and the bonds are shown as S-S. As discussed below, glycosylation in trastuzumab takes place on the asparagine residue Asn (300) (only one of the present glycan structures is shown here).

Trastuzumab is an IgG1 kappa monoclonal antibody and its primary sequence is shown in Figure 3.2. As can be also seen in Figure 3.1 and Figure 3.3, covalent intra- and inter-chain disulfide bonds stabilise the structure. Each heavy chain presents 4 intra-disufide bridges at the cysteine (Cys) sites 22–96, 147–203, 264–324, 370–428 and the two heavy chains are connected with two inter-chain disulfide bridges formed in the hinge region at Cys (229) and Cys (232). Each light chain presents two intra-chain disulfide bonds, at the sites 23–88 and 134–194, and connects to the respective heavy chain with an inter-chain disulfide bridge at L214–H223. In addition to covalent disulfide bonds, non-covalent interactions between pairs of domains hold the chains together, further stabilising the overall structure. Therefore, even if the inter-chain disulfide bonds are broken, the antibody structure may still be maintained. Similarly, each one of the regions is folded into a compact, globular domain—mainly composed

of antiparallel β-sheet secondary structure motifs—linked to one another by short stretches of extended polypeptide chain (Figure 3.3). This domain conformation is, as well, stabilised by the internal disulfide bond and powerful non-covalent forces, which define the secondary structure of the molecule.

Heavy chain – Average Mass = 49284.76 Da

```
1   EVQLV ESGGG LVQPG GSLRL SCAAS GFNIK DTYIH WVRQA PGKGL EWVAR
51  IYPTN GYTRY ADSVK GRFTI SADTS KNTAY LQMNS LRAED TAVYY CSRWG
101 GDGFY AMDYW GQGTL VTVSS ASTKG PSVFP LAPSS KSTSG GTAAL GCLVK
151 DYFPE PVTVS WNSGA LTSGV HTFPA VLQSS GLYSL SSVVT VPSSS LGTQT
201 YICNV NHKPS NTKVD KKVEP KSCDK THTCP PCPAP ELLGG PSVFL FPPKP
251 KDTLM ISRTP EVTCV VVDVS HEDPE VKFNW YVDGV EVHNA KTKPR EEQYN
301 STYRV VSVLT VLHQD WLNGK EYKCK VSNKA LPAPI EKTIS KAKGQ PREPQ
351 VYTLP PSREE MTKNQ VSLTC LVKGF YPSDI AVEWE SNGQP ENNYK TTPPV
401 LDSDG SFFLY SKLTV DKSRW QQGNV FSCSV MHEAL HNHYT QKSLS LSPGK
```

Light chain – Average Mass = 23443.14 Da

```
1   DIQMT QSPSS LSASV GDRVT ITCRA SQDVN TAVAW YQQKP GKAPK LLIYS
51  ASFLY SGVPS RFSGS RSGTD FTLTI SSLQP EDFAT YYCQQ HYTTP PTFGQ
101 GTKVE IKRTV AAPSV FIFPP SDEQL KSGTA SVVCL LNNFY PREAK VQWKV
151 DNALQ SGNSQ ESVTE QDSKD STYSL SSTLT LSKAD YEKHK VYACE VTHQG
201 LSSPV TKSFN RGEC
```

Figure 3.2: Amino acid sequence of the heavy (450 amino acids) and light (214 aa) chain of trastuzumab. The heavy chain is approximately 49 kDa, while the light chain almost 23.5 kDa, resulting in a total mass of ~148 kDa including the glycan structures (Figure 3.6). The cysteine residues that contribute to the formation of disulfide bonds are highlighted in yellow and the glycosylation site is highlighted in orange.

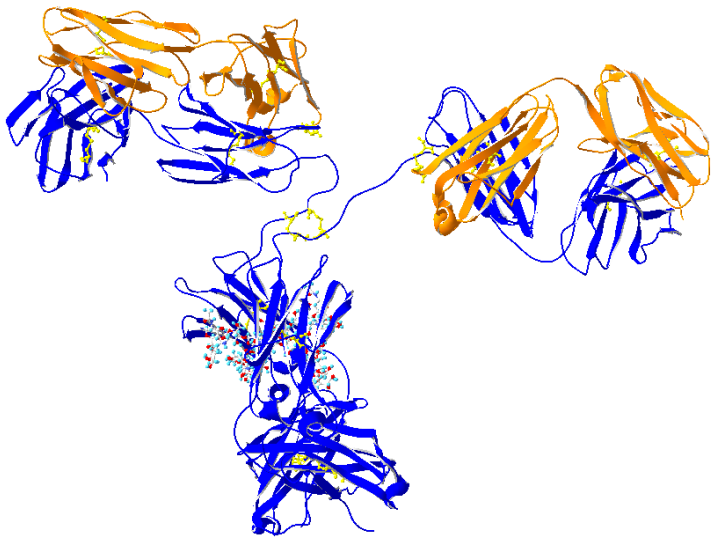


Figure 3.3: Cartoon representation of an IgG2 antibody with PDB ID 1IGT.[30] The heavy chains are shown in blue, light chains are in orange, inter- and intra-chain disulfide bonds are shown in yellow and the attached glycans are shown in ball-and-stick representation. The image is not an accurate representation of trastuzumab, but serves as an example of a complete immunoglobulin structure. Swiss-PDB viewer was used for generating the image.[31]

The antibody-antigen binding is mainly based on steric complementarity between the antibody's CDRs and the antigen's epitope, while a variety of non-covalent interactions (Van der Waals, hydrophobic interactions, hydrogen bonds and ionic interactions) are also involved. The degree of complementarity determines the strength of the binding, and very slight changes in the shape of either can lead to weaker binding or no binding at all. It is therefore of major importance that the antibody maintains its structure after conjugation with the drug-molecule, and the simplest way to investigate such changes is by comparing the secondary structure of the samples.

3.1.2. Toxins

There are several types of drug molecules that are currently used for antibody conjugation. Maytansins (such as DM1 on trastuzumab, used in the commercial product Kadcyla) and auristatins (*e.g.* MMAE on brentuximab, used in Adcetris) act by blocking the polymerisation of tubulin, while calicheamicins target DNA. In the present study, MMAE (monomethyl auristatin E) was the cytotoxic agent attached to trastuzumab. However, in order to achieve stability of the ADC during formulation, storage and in the bloodstream after administration, and efficient release of the pharmacologically active molecule once the ADC enters the tumour cell, various linkers have been designed, which covalently bind the antibody and the cytotoxin.

Linkers are divided in noncleavable and cleavable. As their name declares, noncleavable linkers remain intact during intracellular metabolism and the cytotoxic agent is released after lysosomal degradation of the antibody. Cleavable linkers are metabolised in the cell and release the drug either after hydrolysis (for acid-labile hydrazine linkers) or enzymatic cleavage (peptide-based linkers) or reduction (disulfide linkers). Conjugation of the linker/drug molecule takes place on either the lysine amines (*e.g.* Kadcyla) or the cysteine sulfhydryl groups (*e.g.* Adcetris) of the antibody. Lysines are usually highly available and exposed on the antibody molecule compared to cysteines (*e.g.* for trastuzumab there are 90 lysines and 32 cysteines), and form stable amide bonds after direct reaction with the linkers used. Cysteines are normally involved in the formation of disulfide bridges and therefore partial reduction is a necessary intermediate step in order to form the active sulfhydryl groups. As a result, lysine-based conjugation leads to greater heterogeneity with different drug-to-antibody ratio species, while in the case of cysteine-conjugation the four inter-chain disulfide bonds are typically targeted, producing up to eight possible sites for conjugation after they are partially reduced (Figure 3.4).

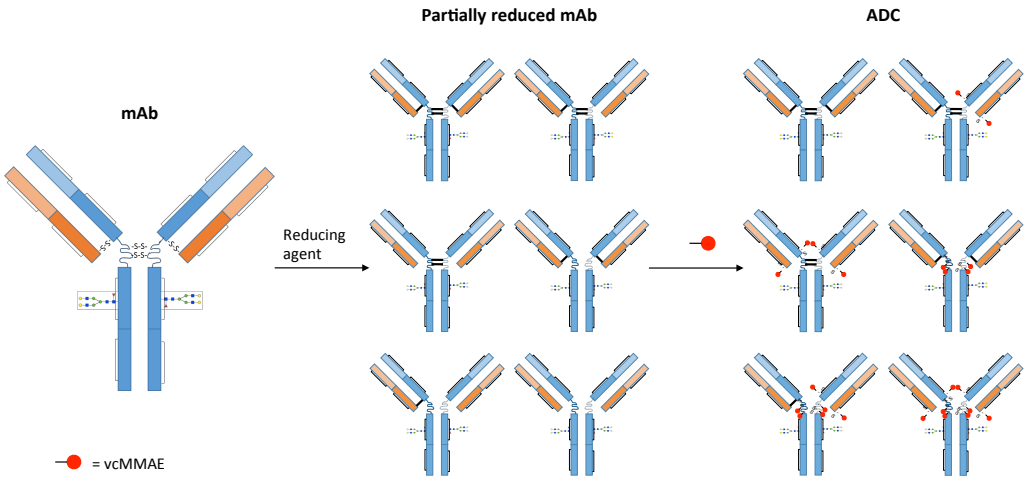


Figure 3.4: Schematic representation of conjugation of the native trastuzumab with the linker/drug vcMMAE. Partial reduction of the inter-chain disulfide bonds leads to six different combinations of the available sulfhydryl groups. Subsequent conjugation can lead to a respective heterogeneous mixture of trastuzumab with the number of attached toxins varying from 0–8.

The whole structure (toxin, spacer and linker) used in the present study is vedotin (vcMMAE). It consists of the toxin MMAE, the PABC spacer (p-aminobenzoylcarbamate) and the valine-citrulline peptide (cleavable linker), while 6-maleimidohexanoic acid is directly attached to the antibody via a cysteine residue (Figure 3.5). The peptide linker of vcMMAE allows easy and fast release of the cytotoxic molecule after hydrolysis of the valine-citrulline peptide by cathepsin B and loss of the PABC molecule. The noncharged MMAE can further diffuse into neighbouring cells exerting its cytotoxicity.[32] The potency of the ADC can be adjusted by regulating the number of the attached cytotoxic molecules. Previously published studies have shown that the best therapeutic window is achieved with 2–4 drug molecules per antibody, while a higher DAR (drug to antibody ratio) might be recognised by the organism as a damaged form and therefore cleared out.[33,34]

Ideally the partial reduction and the subsequent conjugation should be highly controlled processes in order to avoid production of a heterogeneous mixture with varying numbers of cytotoxic molecules per antibody. The targeted residues are usually located on the Fc or constant region, or else interference in antigen-binding might be risked. The conjugation sites and the number of the cytotoxins attached to the antibody should be investigated, so that batch-to-batch reproducibility can be confirmed.

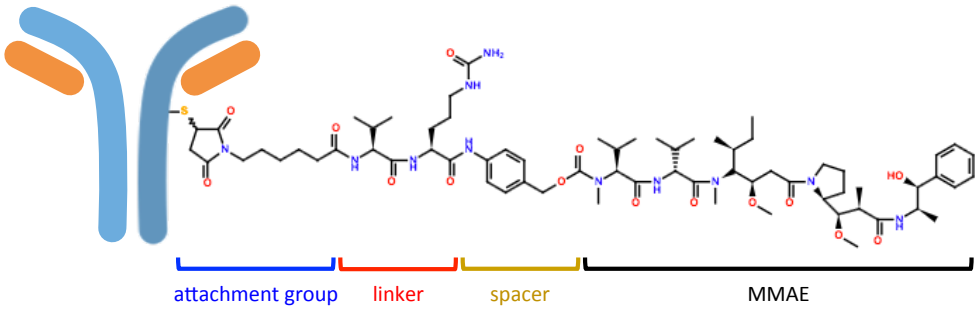


Figure 3.5: Monoclonal antibody linked to the vcMMAE structure via 6-maleimidohexanoic acid (attachment group). Vedotin (vcMMAE) consists of the toxin MMAE (monomethyl auristatin E), a spacer (PABC) and the peptide-based linker valine-citrulline. 6-maleimidohexanoic acid is attached to the cysteines of the antibody. The vcMMAE structure was sketched in chemspider.com (http://www.chemspider.com/StructureSearch.aspx).

3.1.3. Glycosylation

One of the most important post-translational modifications on antibodies is glycosylation. Although glycosylation is a very consistent and reproducible process under physiological conditions, production of glycotherapeutic agents is characterised by great heterogeneity depending on the culture conditions and usually results in protein preparations that contain more than one glycoforms, which as a consequence can affect the potency and safety of the product. The current regulatory guidelines, therefore, require that the manufacturers should assure that a consistent human glycoform profile is maintained for recombinant mAbs and among different lots.

The most common types of glycosylation are the N- and O-glycans. N-type is linked to the amino group of asparagine, when the sequence pattern Asn-X-Ser or Asn-X-Thr is met, with X being any amino acid except proline. O-glycans are attached to the hydroxyl group of serine or threonine. IgG monoclonal antibodies have at least one N-glycosylation site located at asparagine 297 of each heavy-chain C_H2 domain of the Fc portion,[35] while the glycosylation site of recombinant IgG isotypes might vary slightly, depending on the manufacturing process. Such is the case of trastuzumab, in which the glycosylation site is located on Asn (300). In eukaryotes, protein glycosylation takes place in the endoplasmic reticulum (ER) and the Golgi apparatus. During the initial stage, which is common for all proteins, a block of 3 glucoses (Glc), 9 mannoses (Man) and 2 N-acetylglucosamines (GlcNAc) is attached to the protein ($Glc_3Man_9GlcNAc_2$) in the ER. A cycle of addition and removal of glycans is achieved with the help of transferases and glycosidases, until the protein is correctly folded.[36] Further processing follows in the Golgi apparatus, which aims to diversify the glycans, leading to the final glycoforms.[37]

In the case of IgG antibodies, sugars usually belong to the bi-antennary complex type, presenting a conserved heptasaccharide core, which is composed of two N-acetylglucosamine (GlcNAc), three mannose (Man) and two more GlcNAc residues. Additional residues might be present, such as fucose, galactose, and N-acetylneuraminic acid (NANA). Despite the considerable heterogeneity,[38] the most abundant structures of N-linked glycans in mAbs—also observed in the present study—are usually G0F, G1F and G2F, according to the number of galactose residues on the two branches (Figure 3.6). Due to their position in the hydrophobic surface of the C_H2 domain of the antibody molecule (Figure 3.3) and their subsequent noncovalent interactions with the protein surface, glycans play a major role in protein conformation and stabilisation, maintaining and regulating binding to the Fc receptors.[39,40] Deviation in the glycosylation of immunoglobulins is linked to several diseases (such as rheumatoid arthritis),[41] while the presence of certain saccharides (*e.g.* α1,3-fucose) is related with high immunogenicity.[42] Conclusively, both the present glycoforms and the glycosylation sites need to be well characterised in mAbs, confirming consistency in glycan composition among different samples and providing an additional evidence of the product's stability and potency and the manufacturer's control over the whole process. In the case of trastuzumab, G0F, G1F and G2F were observed, accounting for ~1.5% of the antibody's mass.

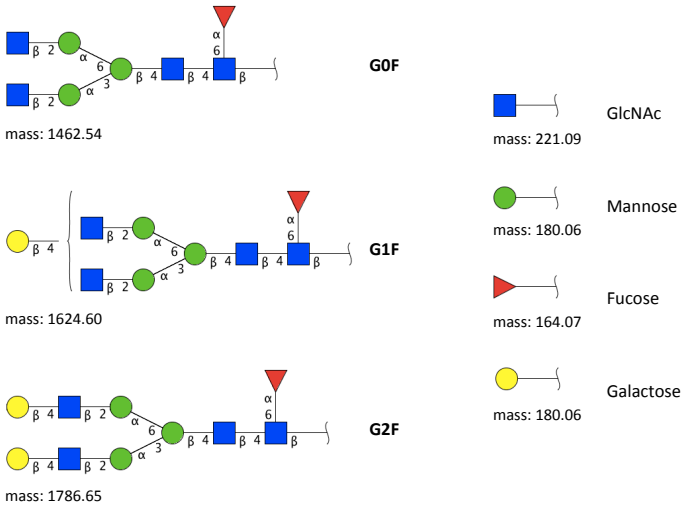


Figure 3.6: Structures of the three most abundant N-glycans found in recombinant antibodies. The monoisotopic* masses are given for the free non-conjugated monosaccharides and for the non-protein bound oligosaccharides G0F, G1F and G2F. The respective masses are 18 Da lower within the

* The monoisotopic mass is calculated based on contributions only from the lightest isotope for each element.

glycopeptide as shown in Figure 3.11. The representation follows the symbol nomenclature outlined by the Consortium for Functional Glycomics (http://www.functionalglycomics.org/static/consortium/Nomenclature.shtml).

3.1.4. Deamidation

Deamidation is one of the most common amino acid modifications and can be observed on asparagine (N) and less frequently on glutamine residues of a protein sequence. Previous studies have shown the heavy chain N255 residue as one of the modification hotspots observed on the trastuzumab molecule,[43,44] while N387 has also been shown to be susceptible.[45] Deamidation is usually favoured on asparagine residues that are solvent accessible and are often found in flexible regions of the molecule, while the most common case is when glycine (G) is the following amino acid as its short side chain further increases accessibility to the asparagine side chain. Asparagine sites followed by amino acids with larger side chains are somewhat protected due to steric hindrance. Exposure to high pH (~8) and temperature can further promote deamidation. The focus in this study was mainly on the asparagine residues, which constitute 3.76% of the sequence (19 residues on HC and 6 on LC), and more specifically in cases where these are followed by glycine. In the trastuzumab sequence three NG sites can be found (N55, N318 and N387 on HC) and deamidation was observed on all three of them (see MS results, Section 3.3.1).

The mechanism of deamidation is shown in Figure 3.7. The reaction proceeds via formation of an L-succinimide intermediate with loss of ammonia (17 Da), which is in equilibrium with its D-enantiomeric form. The succinimide intermediate was present in most of the cases that deamidation was observed in the present study. Hydrolysis (gain of 18 Da) of any of the enantiomers, results in a mixture of aspartate (D) and isoaspartate, with the last leading to a significant change in the backbone of the protein by introducing a methylene. As a consequence, deamidation causes charge heterogeneity in proteins and often results in reduction or loss of biological activity,[46] with some exceptions where the protein remains unaffected.[47] Isomerisation of aspartate to isoaspartate has been found to play a role in aging, autoimmune response, cancer and neurodegeneration (such as Alzheimer's disease),[48] while in the case of recombinant monoclonal antibodies it can potentially affect the clinical efficacy depending on the location. The importance of identifying the sites and amount of deamidation is therefore crucial, in order to evaluate the possible outcomes. However, the fact that aspartic and isoaspartic acid have identical masses and similar chemical properties (with isoaspartate eluting only slightly earlier), makes it hard to differentiate between the two unless specific experiments are conducted.[48,49]

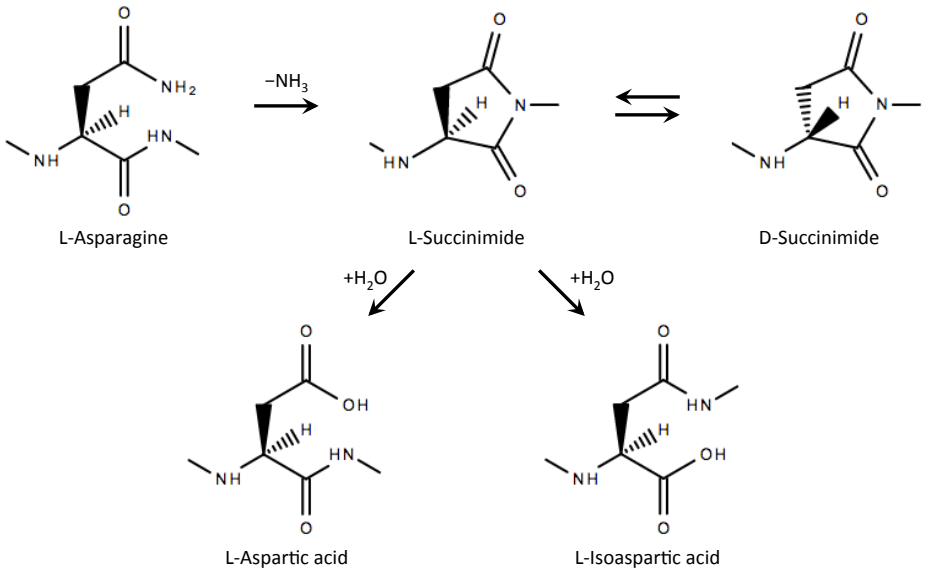


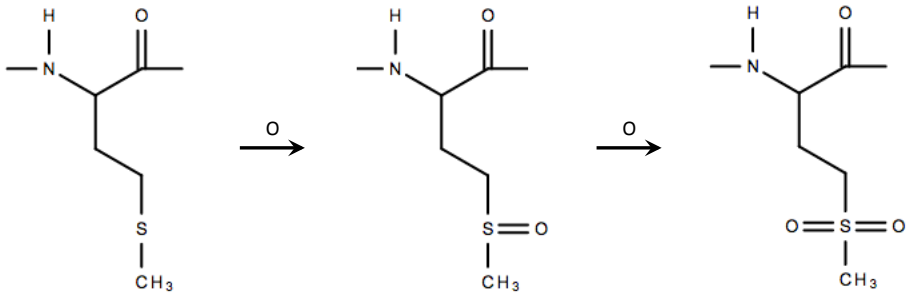
Figure 3.7: Mechanism of deamidation reaction on asparagine. Formation of the cyclic L-succinimide with loss of ammonia and subsequent hydrolysis to aspartic or isoaspartic acid happens, resulting in a total mass increase of 1 Da for the modified peptide. The structures were sketched in chemspider.com (http://www.chemspider.com/StructureSearch.aspx).

3.1.5. Oxidation

Oxidation is the protein modification that results in the addition of one or more oxygen atoms to a susceptible side chain, after reacting with atmospheric oxygen or oxygen radicals in solution. On monoclonal antibodies, it is most frequently observed on methionine and tryptophan residues,[50-52] it has been reported on histidine under light stress (photooxidation),[53] while unpaired cysteines can also be potential sites of oxidation.[54] The hotspot for methionine oxidation on trastuzumab has been previously reported to be the M255 residue.[52,55] Since methionine is one of the most hydrophobic amino acids, exposure to aqueous environments can promote modification. As a result, the presence of polar amino acids in the surrounding environment of a methionine residue favours oxidation on the specific site.[56] Like most proteins, trastuzumab has a low content of methionine residues (0.90%) with five of them located on the heavy chain and one on the light chain, while oxidation was observed on two of them (one on the heavy and one on the light chain).

Figure 3.8 shows the steps of oxidation for methionine. Addition of one atom of oxygen leads to methionine sulfoxide (Mso), increasing the molecular mass of the peptide by 16 Da, while further oxidation results in the addition of an extra oxygen and the formation of methionine sulfone (Msn) (increase of 32 Da). Methionine oxidation has been found to decrease structural stability and accelerate aggregation,[57] it can cause

protein misfolding, which disrupts the function,[58] while in the case of antibodies it has been found to affect interaction with receptors and binding.[59] As with deamidation, it is therefore important to identify the sites and the extent to which oxidation occurs in order to assess the possible consequences on potency and immunogenicity. In the present study, only methionine sulfoxide was observed, mainly on M255 (heavy chain), and an insignificant amount on M4 (≤ 1%, on the light chain).



Methionine | Methionine sulfoxide (Mso) | Methionine sulfone (Msn)

Figure 3.8: Oxidation steps on methionine. Addition of one oxygen atom results in the formation of methionine sulfoxide (Mso), while the second oxygen atom leads to methionine sulfone (Msn) and a total molecular mass increase of 32 Da.

3.2. Materials and Methods

3.2.1. Materials

L-Histidine (H8000), L-Histidine monohydrochloride monohydrate (H5659), Tween 20 (P9416) and D-(+)-trehalose dihydrate (T9531) were purchased from Sigma-Aldrich. Ultrapure water was used (18 MΩ·cm at 25 °C) in all cases. The latex spheres (3100A) were from ThermoFisher Scientific. For the CD calibration a Neodymium and Holmium filter from Starna were used. The ammonium D-10-camphorsulfonate was from Katayama chemical Co. 05-1251, ammonia solution 35% w/w was purchased from Scientific Laboratory Supplies (CHE1136), benzene from Aldrich (401765), and the Na[Co(S,S-EDDS)], Na[Co(R,R-EDDS)] standards (CD0038) were synthesized at the University of Warwick.[60,61] Potassium chromate was purchased from Aldrich (216615) and potassium hydroxide pellets AnalaR from BDH (102104V). All samples—the native monoclonal antibody prior to conjugation (Herceptin/Trastuzumab), the ADC (Her vcE) made using the traditional solution method and the ADC made using ADC Biotechnology's proprietary lock-release method—were kindly provided by ADC Biotechnology as follows:

Table 3.1: The native monoclonal antibody (Herceptin) and the antibody drug conjugate (Her vcE) samples as received by ADC Biotechnology. vcE is the conjugated toxin used as the anticancer drug (MMAE) with the linking to the antibody structure (altogether named as vedotin or vcMMAE), and Formulation A is a solution of 5 mM histidine-HCl buffer, 50 mM trehalose, 0.01% w/v Tween 20 and pH 6.0. The names and the colours (blue, green, red, orange) of the samples (left column) are according to the figures in the results section.

Sample name in text	Sample ID	Sample Reference No	Protein Concentration (mg/mL)	Volume supplied (mL)	Formulation
mAb (S)	Herceptin (Solution)	038 151	1.11	5 x 1 mL	A
mAb (LR)	Herceptin (Lock-Release)	053 113	1.07	5 x 1 mL	A
ADC (S)	Her vcE (Solution)	045 016	1.09	5 x 1 mL	A
ADC (LR)	Her vcE (Lock-Release)	045 007	0.98	5 x 1 mL	A

3.2.2. Spectroscopic data collection

Peptide mapping analysis was performed using an Acquity ultra-performance liquid chromatography (UPLC) PDA (Photodiode array) detector (Waters) coupled to a Xevo G2 Q-TOF (quadrupole time-of-flight) mass spectrometer with ES and Nanospray (Waters). The analytical column was an Acquity UPLC BEH C18, 1.7µm, 1.0 x 50 mm. Mass spectra were processed with MassLynx 4.1, and BiopharmaLynx was used

for confirmation and to aid further interpretation. The light and heavy chain sequences of trastuzumab were entered in each software, where the masses of the tryptic digestion-generated peptides were computed allowing for up to 5 missed cleavages. Carboxymethylation of cysteine residues was set as a fixed modification, while lysine clipping at the C-terminus was set as variable modification. Identification of the glycopeptides—yielding the glycan and peptide structures—and the glycosylation site was allowed with BiopharmaLynx. The data were also manually searched for methionine oxidation, asparagine deamidation and the succinimidyl intermediate. The areas of the modified and unmodified peptides were calculated by integration of the respective peaks in the extracted chromatograms. The amount of modification on each site is expressed as a percentage of the total sum of the modified and unmodified areas.

The ***quantitative LC-MS/MS*** analysis was performed using an Acquity UPLC PDA detector (Waters) in-line with a Xevo TQ-S Triple Quadrupole tandem mass spectrometer with ES (Waters). The analytical column was an Acquity UPLC BEH C18, 1.7µm. For each of the conjugated samples (ADC (S) and ADC (LR)), sample treatment—with the necessary reactions prior to MS analysis—was conducted in triplicate. Eight replicate measurements in total were acquired for each conjugated sample, from which two were excluded in each case. All data were analysed using MassLynx 4.1, and the quantification was performed using the integrated areas of the peaks of interest in the MRM plot (multiple reaction monitoring plot). The mass fragment corresponding to m/z 359.20 was used as the marker peptide, and the ions with m/z 686.50 and m/z 718.50 were fragments resulting from the drug molecule of the drug-conjugated peptides of the light (LC) and heavy (HC) chain as shown in Table 3.4. The sum of the integrated areas from the two transitions for each drug-conjugated peptide was normalised by dividing by the integrated area of the marker peptide.

The ***DLS*** and ***CD data*** were collected on the instruments as mentioned in Chapter 2 (Section 2.2.2). The data were collected in six replicates and processed as discussed in Chapter 2, with a few differences that are described below. The high concentration (1 mg/mL) CD data were collected on a JASCO J-815 spectropolarimeter.

For the ***DLS data*** 6 measurements at each temperature were averaged for each replicate. For each sample, the data are presented in terms of intensity and number values as an overall mean of the averaged values over the 6 replicates. The standard deviation was calculated over six values in each case, using the averaged measurements of each replicate.

For the ***CD spectra***, the same spectrum of the formulation was subtracted from each replicate in the case of the first three replicates, while for the last three replicates spectra of the formulation were recorded on the same day and cuvette as their respective samples and used for the baseline-correction. The data were zeroed by subtracting the averaged value from 250–240 nm from the whole spectrum (mdeg) and converted to Δε units. The data were averaged over the 6 replicates (as presented in the text) and the wavelength step was changed to 1 nm. For the secondary structure analysis the truncated spectra and SSNN were used.[62] For the high concentration (1 mg/mL) CD spectra, 5 replicates were recorded for each sample and baseline-corrected by subtracting the same formulation spectrum. Zeroing, conversion to Δε units and averaging followed, and the data were shifted by −3 nm to correct for the difference in the calibration between the two instruments.

The ***balance, pH meter, refractometer***, ***centrifuges*** and ***water purification*** system were as described in the spectroscopic data collection of Chapter 2 (Section 2.2.2).

3.2.3. Measurements

For the ***peptide mapping experiments,*** the peptide mixture was infused onto the column at a flow rate of 10 µL/min and a total volume of 15 µL was injected. The column temperature was 50.0 °C, elution was performed at 0.350 mL/min and a gradient as shown in Table 3.2, and electrospray ionisation (ESI) in the positive ion mode was used. The spectrometer was calibrated (for mass correction) prior to analysis using [Glu1]-Fibrinopeptide B. MS spectra were recorded over an m/z range of 50-2000 with a scan duration of 0.5 sec, capillary voltage was 2.5 kV, cone voltage 40.0 V, source temperature 120 °C and desolvation temperature was 400 °C. The cone gas flow was set to 50 L/h and the desolvation gas flow was 550 L/h.

Table 3.2: Liquid chromatography gradient table for the peptide mapping Q-TOF experiments. Solvent A: 0.1% formic acid in water, solvent B: 0.1% formic acid in acetonitrile. The total run time was 45 min.

Time (min)	Flow rate (mL/min)	%A	%B	Curve
Initial	0.350	100.0	0.0	Initial
0.30	0.350	100.0	0.0	6
34.00	0.350	55.0	45.0	6
35.00	0.350	15.0	85.0	6
37.00	0.350	15.0	85.0	6
38.00	0.350	100.0	0.0	6
39.00	0.350	0.0	100.0	6
40.00	0.350	100.0	0.0	6
41.00	0.350	0.0	100.0	6
42.00	0.350	100.0	0.0	6
43.00	0.350	0.0	100.0	6
44.00	0.350	100.0	0.0	6
45.00	0.350	100.0	0.0	6

For the ***mass spectrometric quantitation,*** the peptide mixture was infused onto the column at a flow rate of 5 µL/min and a total volume of 5 µL was injected. The column temperature was 60.0 °C, elution was performed at 0.400 mL/min and a gradient as shown in Table 3.3. The mass spectrometer was operated in the positive ion electrospray ionisation (ESI) mode and calibrated prior to analysis using 0.1% ortho-phosphoric acid. The capillary voltage, cone voltage and collision energy were set to 2.4 kV, 30.0 V and 25.0 eV, respectively. The source temperature was set to 150 °C and the desolvation temperature was 550 °C with a desolvation gas flow of 1100 L/h. The MRM (multiple reaction monitoring) acquisition mode was used, following the parent-to-daughter ion transitions for a marker peptide and the peptides attached to the drug molecule as shown in Table 3.4.

Table 3.3: Liquid chromatography gradient table for the mass spectrometric quantitative analysis on the TQ-S spectrometer. Solvent A: water, solvent B: acetonitrile. The total run time was 45 min.

Time (min)	Flow rate (mL/min)	%A	%B	Curve
Initial	0.400	100.0	0.0	Initial
0.50	0.400	100.0	0.0	6
37.50	0.400	55.0	45.0	6
40.00	0.400	3.0	97.0	6
41.00	0.400	3.0	97.0	6
41.50	0.400	100.0	0.0	6
42.00	0.400	0.0	100.0	6
42.50	0.400	100.0	0.0	6
43.00	0.400	0.0	100.0	6
43.50	0.400	100.0	0.0	6
44.00	0.400	0.0	100.0	6
44.50	0.400	100.0	0.0	6
45.00	0.400	100.0	0.0	6

Table 3.4: Transitions of parent to fragment ions monitored in the MRM mode for the quantitative analysis of drug molecules in the antibody structure. The marker peptide, the drug peptides and the retention time (RT) at which they are eluted are shown. MMAE is the drug molecule without the linker and monoisotopic mass 718.50 Da. LC: light chain, HC: heavy chain.

Transition (m/z)	RT (min)	Peptide (parent ion)	Fragment ion	Peptide property
886.99 > 359.20	22.24 ± 0.16	LC46–61	y''3	Marker
812.44 > 686.50	32.66 ± 0.17	LC212–214	MMAE fragment	Drug
> 718.50			MMAE	Drug
884.48 > 686.50	30.27 ± 0.18	HC222–225	MMAE fragment	Drug
> 718.50			MMAE	Drug

3.2.4. Temperature dependent measurements

For all ***CD temperature-dependent experiments***, the same parameters were used: wavelength range 250–180 nm, 1 s data integration time (DIT), 2 nm band width, data pitch 0.2 nm, 100 nm/min continuous scanning speed. Data were obtained every 1 °C from 20 °C to 100 °C, at a ramp rate of 1.0 °C/min. 30 s were allowed as equilibration time once the desired temperature was kept within ± 0.1 °C for 5 seconds. For the last three replicates, CD spectra were also recorded at the conclusion of the temperature experiment, at least 30 min after the samples had been cooled down to 20 °C. Each spectrum is the average of 4 accumulations, while 16 accumulations spectra were collected for each sample before the temperature experiment, at 20 °C. The total duration time of the temperature experiment was ~ 8 h. A 0.1 cm path length cuvette

was used for ~0.1 mg/mL trastuzumab solutions. For the high concentration solutions (~1 mg/mL) and for the excipients of the formulation a demountable cuvette with a nominal path length of 0.01 cm was used. The path length of the cuvette was measured by using the absorbance of 0.02 M potassium chromate and the Beer-Lambert Law with an extinction co-efficient of 4830 $M^{-1}cm^{-1}$ at 372 nm, and was determined to be 0.01071 ± 0.00003 cm. The value 0.01071 was used in the analysis. The J-1500 CD spectrophotometer was calibrated according to SOP 48.3 (attached in Appendix C) at the beginning and at the end of the temperature-dependent experiments, in order to confirm that its operation was within the accepted criteria (see below, CD calibration, Section 3.2.6).

For the ***DLS experiments***, the viscosity was estimated[63] for the used formulation (Table 3.5), by running measurements of solutions consisting of 1 µL of polystyrene latex spheres of known size (100 nm) in 99 µL of blank solution at each temperature. The Z-average[64] sizes were used, as the blank was monodisperse (polydispersity index < 0.1). In a temperature range from 20–75 °C with a 5 °C temperature interval, 6 measurements were recorded for each temperature. 300 s were allowed as the equilibration time after the desired temperature was reached and each measurement consisted of 10 runs, with 200 s total duration. Six replicates were recorded for each one of the four samples.

Table 3.5: Corrected values for viscosity at each temperature and refractive index for the used blank solution (Formulation A).

Temperature (°C)	20	25	30	35	40	45	50	55	60	65	70	75	Refractive Index
Formulation A	1.029	0.909	0.805	0.717	0.652	0.601	0.555	0.515	0.476	0.440	0.409	0.379	1.335

3.2.5. Sample preparation

All protocols for the reactions, sample treatment and standard operating procedures (SOPs) used for the ***MS experiments*** are proprietary of BioPharmaSpec and therefore cannot be published herein.

Formulation A was prepared as a blank for the CD and DLS experiments. A 30 mM histidine buffer was prepared by mixing 530 µL of 60 mM L-histidine monohydrochloride solution with 470 µL of 60 mM L-histidine free base solution and 1 mL of water. The 30 mM Histidine-HCl buffer was mixed with 0.2% w/v Tween 20, 100 mM trehalose and water in suitable ratios, resulting in a final formulation of 5 mM histidine buffer, 50 mM trehalose, 0.01% w/v Tween 20 and pH 6.01.

For the ***CD temperature dependent measurements*** the samples and the formulation were diluted by a factor of 10 with water. CD data of the samples in their original concentration (~1 mg/mL) were also recorded at 20 °C, using a 0.01 cm cuvette.

For the ***DLS experiments*** samples and formulation (blank) were used in their original concentrations. The formulation solution with the latex spheres was sonicated for 2–3 min, and all solutions (samples and formulation) were filtered through 0.2 μm syringe filters (Thermo Scientific) directly into the Quartz cuvette prior to the measurement. The refractive index of the blank was measured (Table 3.5).

3.2.6. CD calibration

The target values and the measured values from the CD spectrophotometer calibration before and after the temperature-dependent experiments are shown in Table 3.6 and Table 3.7 for each one of the performed tests. For the calibration of the wavelength accuracy, the neodymium and holmium filters, ammonia[65,66] and benzene[67,68] vapour were used (Table 3.6). Ammonia vapour was used only at the end of the experiments. To determine the scaling factor for the CD intensity, a 0.06% (w/v) aqueous ammonium D-10-camphorsulfonate (ACS)[69] solution and the enantiomeric forms (*R,R*)- and (*S,S*)-Na[Co^{III}(EDDS)] (ethylenediaminedisuccinate)[61,70] were used (Table 3.7).

The measured values from the wavelength accuracy tests (Table 3.6) are within the accepted limits in the region 240–640 nm for both calibrations—before and after the experiments. At lower wavelengths (194-209 nm), the accuracy presents a deviation of ~−0.3 nm from the accepted values. From the intensity performance tests (Table 3.7), the maximum of the ACS solution peak at 291 nm appeared to be 5.3–5.6% lower than the target value, while the deviation for the (*R,R*)-Na[Co^{III}(EDDS)] enantiomer was 3.0–4.6% at 211.5 nm. The maxima for the two enantiomeric forms appeared at 210.5 nm rather than at 211.5 nm, which is in agreement with the noticed wavelength deviation at this region. The ratio of the signals at 193 nm and 291 nm for the ACS solution was calculated as 2.02 and 2.06 before and after the experiments respectively, with the last value meeting the acceptance criteria, as this is normally expected to be 2.05-2.10.

The observed wavelength deviation and the intensity variation at the far-UV region—where the protein spectra are analysed—could potentially affect the secondary structure prediction and the spectral fit with the used algorithm. In order to adjust for these differences the spectra can be artificially shifted by +1 nm and a scaling factor of

~1.04 can be applied for the intensity.[71,72] However, no effect from the calibration was observed on the secondary structure prediction, and therefore the corrections were considered unnecessary.

Table 3.6: Target wavelengths for the HT maxima for each one of the calibration tests, measured wavelengths at the beginning and at the end of the experiments and the respective deviations are presented. The measured values should be ± 0.8 nm from the target values. Values that do not meet the criterion are noted in red.

Test	Target Wavelength (nm)	Wavelength before experiment (nm)	Deviation (nm)	Wavelength after experiment (nm)	Deviation (nm)
Neodymium filter	586.0	585.5	−0.5	585.4	−0.6
Holmium filter	637.5	638.4	0.9	638.2	0.7
	536.6	536.2	−0.4	536.0	−0.6
	460.2	459.6	−0.6	459.6	−0.6
	453.7	453.4	−0.3	453.2	−0.5
	445.7	445.2	−0.5	445.2	−0.5
	360.9	360.4	−0.5	360.2	−0.7
	333.9	333.4	−0.5	333.4	−0.5
	287.6	287.2	−0.4	287.2	−0.4
	279.4	279.0	−0.4	278.8	−0.6
Benzene vapour	259.0	258.7	−0.3	258.6	−0.4
	253.0	252.6	−0.4	252.6	−0.4
	247.2	246.8	−0.4	246.8	−0.4
	241.7	241.3	−0.4	241.2	−0.5
Ammonia vapour	208.6	-	-	207.6	−1.0
	204.8	-	-	203.7	−1.1
	201.1	-	-	200.0	−1.1
	197.5	-	-	196.4	−1.1
	194.0	-	-	193.1	−0.9
	190.6	-	-	189.8	−0.8
	187.2	-	-	186.4	−0.8

Table 3.7: The maximum target intensity values at the specific wavelengths and the measured intensities before and after the temperature experiments are shown for the performed tests. The measured values should be ± 1 mdeg from the target values. Values that do not meet the criterion are noted in red.

Test	Wavelength (nm)	Target Intensity (mdeg)	Intensity before experiment (mdeg)	Intensity after experiment (mdeg)
ACS solution (1 cm cuvette)	291	188.0	179.9	180.6
ACS solution (1 mm cuvette)	291	18.8	18.5	18.4
	193	−39.1	−37.5	−37.9
(*R*,*R*)-Na[Co^{III}(EDDS)]	211.5	−50.0	−48.5	−47.7
(*S*,*S*)-Na[Co^{III}(EDDS)]	211.5	50.0	49.3	49.5

3.3. MASS SPECTROMETRY

Mass spectrometry has been widely used for the characterisation of biotherapeutic proteins[73-75] and for distinguishing the lot-to-lot heterogeneity in glycosylation profiles,[76] and is one of the first choice techniques required by the ICH guidelines. Acquiring data of the intact structure is usually the initial step. However, in order to qualitatively identify and quantify any modifications in the primary structure, large proteins should be cleaved into smaller peptides, which will facilitate the analysis within the mass resolution limits.

3.3.1. Q-TOF results

Protein cleavage is accomplished with the use of sequence specific enzymes (proteases), such as Trypsin, Chymotrypsin and Glu-C, either in separate or sequential digestions when improved sequence coverage is required. In the present study, tryptic digestion was applied on trastuzumab resulting in sufficient coverage of 94.90% for the heavy chain and 100.00% for the light chain (see overall coverage map in Appendix A, Figure A1).

Trypsin cleaves at the C-terminal end of lysine (K) and arginine (R) (primary cleavage), while at later stages of the proteolytic process it presents chymotryptic activity and also cleaves at the C-terminal end of tyrosine (Y), phenylalanine (F) and tryptophan (W) (secondary cleavage). A few steps usually precede the tryptic digestion: denaturation is applied in order to make the peptide bonds more accessible to the enzyme, the disulfide bonds are reduced and the free cysteines are protected by carboxymethylation. The resulting peptide mixture from the protein digestion is separated by reversed-phase UPLC where the elution time of each peptide fragment depends on its hydrophobicity, therefore providing a "fingerprint" chromatographic pattern (TIC plot, total ion current) characteristic of the protein under analysis. In conjunction with a mass spectrometer, the effluent enters the electrospray system allowing identification of the expected peptides and modified—due to post-translational modifications (PTMs) or degradation—peptides.

The **peptide mapping** tables for each sample can be found in Appendix A (Table A1–Table A4), where apart from primary and secondary cleavage, missed cleavage sites were also observed (trypsin cleaves at the second or third K/R in turn, *e.g.* peptide LC36–45). The TIC plots (Figure 3.9) for the four samples look quite similar on a first inspection, confirming consistency in identity, homogeneity and

purity of the samples. The most prominent difference appears at RT 23.0–27.0 min, with the two ADCs presenting some peaks due to elution of the drug-conjugated peptides in this region. A closer look at the overlaid plots (Figure 3.10) discloses some more variations, namely at RT 15.93 min the two native antibodies (mAb (S) and (LR)) differ significantly from the conjugated samples due to elution of the non-conjugated version of the peptide HC222–244 at that time. At RT 16.52 min the peptide HC305–323 elutes, and the observed differences are probably due to the varying amounts of deamidation on the asparagine residue N318 (see below for deamidation). Intensity differences in the chromatographic patterns can also result from variations in the tryptic digestion that might occur from sample to sample or varying response to MS. A common heavy chain C-terminal modification, **lysine-clipping**, was observed in three of the samples on the HC443–450 peptide.

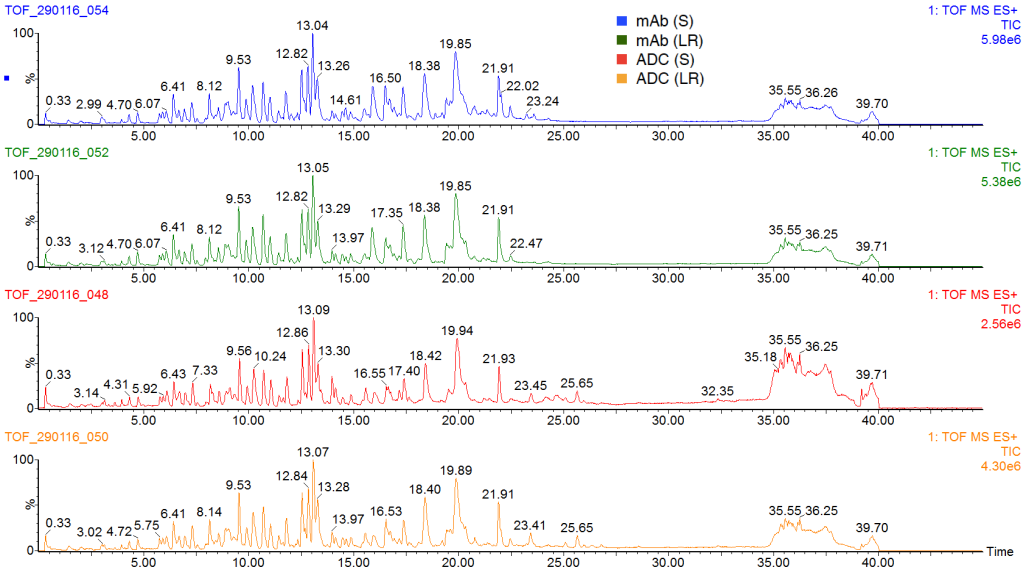


Figure 3.9: TIC (total ion current) plots for the four samples resulting from digestion with trypsin. The chromatographic patterns were obtained by reversed-phase liquid chromatography.

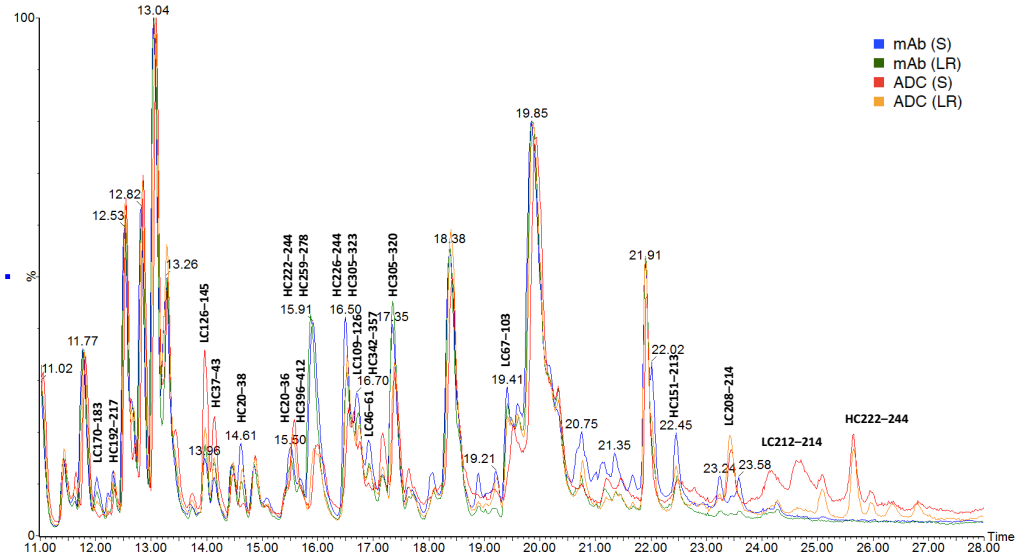


Figure 3.10: Overlaid TIC (total ion current) plots for the four samples resulting from digestion with trypsin. The chromatographic patterns were obtained by reversed-phase liquid chromatography. The retention times (RT) at which the samples differ are indicated with the eluted peptides assigned.

After confirming the primary structure for all four samples, the **glycan structures** and the site of glycosylation were identified. The spectrum of the ADC (S) sample in Figure 3.11 proves the presence of three different glycans (G0F, G1F and G2F) attached to the peptide. The respective peaks were found in the second charge state at m/z 1317.51, 1398.55 and 1479.58, while smaller glycan fragments were also observed and G0F as a detached fragment. The same peaks were found for all four samples (Appendix A, Figure A2) confirming consistency. Figure 3.12 presents the chromatogram peaks, at RT ~4.30 min, corresponding to the three different forms of the glycosylated peptide for the four samples. The similarity of the peaks (G0F, G1F and G2F glycosylation) and the fact that all three forms are observed in the same spectrum is an indication that the glycan structures exist as parental forms of glycosylation, rather than G0F or G1F occurred as fragments from either G1F/G2F or G2F respectively due to loss of galactose residues. The significantly lower intensity of the peak corresponding to the G2F glycosylated peptide—consistent among the four samples—declares the different populations of each structure (inset tables in Figure 3.12). However, a separate analysis would be required in order to confirm the exact populations, which was not necessary in this study as the intention was to obtain an overall image and to confirm the absence of immunogenic glycan structures.

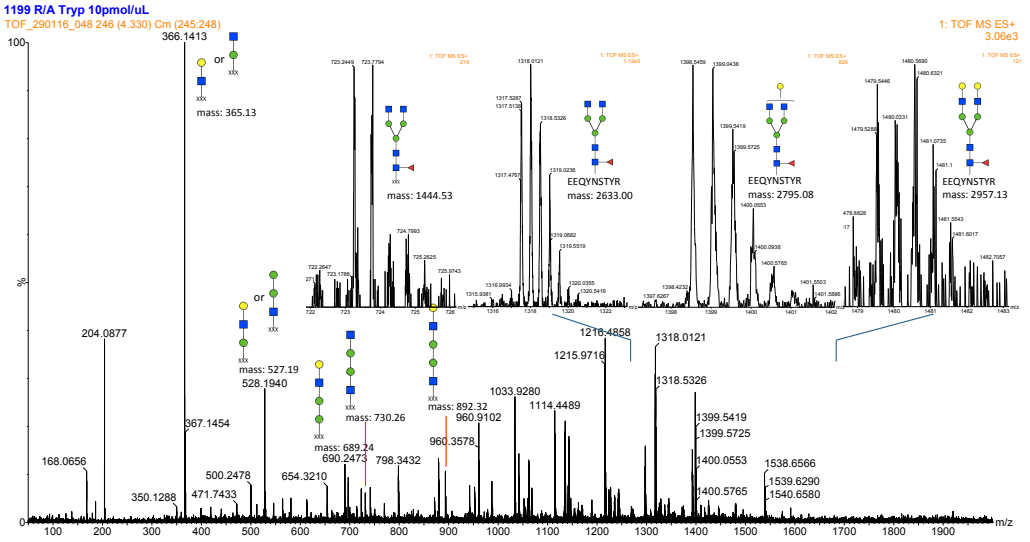


Figure 3.11: The peaks for the glycosylated peptide HC296–304 are shown for the ADC (S) sample. The glycan structures of G0F, G1F and G2F were found attached to the peptide at the $[M+2H]^{2+}$ state with m/z 1317.51, 1398.55 and 1479.58, respectively. G0F is also present as a fragment (m/z 723.26) as well as smaller glycan fragments at the lower end of the spectrum (m/z 366.14, 528.19, 690.25, 731.26, 893.32) as the assigned structures indicate. A comparison of the respective spectra for all four samples can be found in Appendix A (Figure A2).

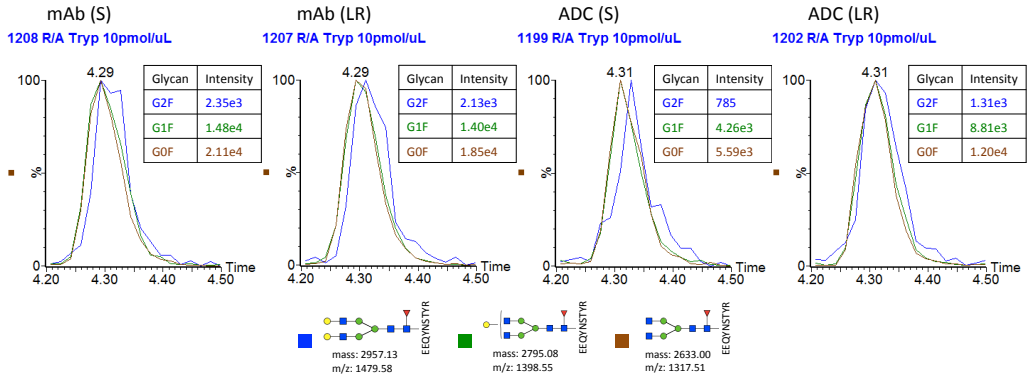


Figure 3.12: Chromatogram peaks corresponding to the glycosylated peptide HC296–304 attached to G2F (blue), G1F (green) and G0F (brown) for each one of the four samples. The G2F peak was shifted by −0.02 min in order to align with the other peaks. The similar shape of the three peaks indicates that all three glycan structures are observed as parental glycosylation forms. The respective intensities indicate different populations of the three structures.

All cysteine residues were assumed to be carboxymethylated and the respective peptides were identified as such. The same peptides were also searched for the cysteine residues on which **conjugation** occurred. The total masses of the peptides with the unmodified cysteines and the vcMMAE molecule were calculated and identified, while the presence of vcMMAE signature fragment ions further confirmed the conjugation sites. The two most abundant signature fragment ions in our spectra had m/z 506.36 and 718.51 (MMAE molecule) and the cleavage sites are shown in Figure 3.13. The conjugation sites were revealed to be on the cysteines that contribute

to the inter-chain disulfide bond between the heavy and the light chain, namely on Cys (223) and Cys (214) respectively (Figure 3.14).

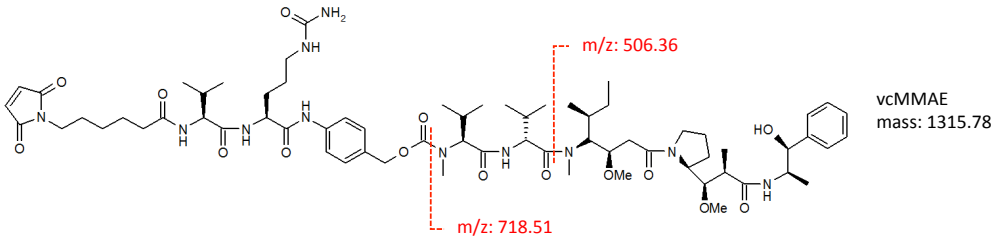


Figure 3.13: The two most abundant signature fragment ions for vcMMAE linker-drug that were observed in the under-analysis samples (following figures).

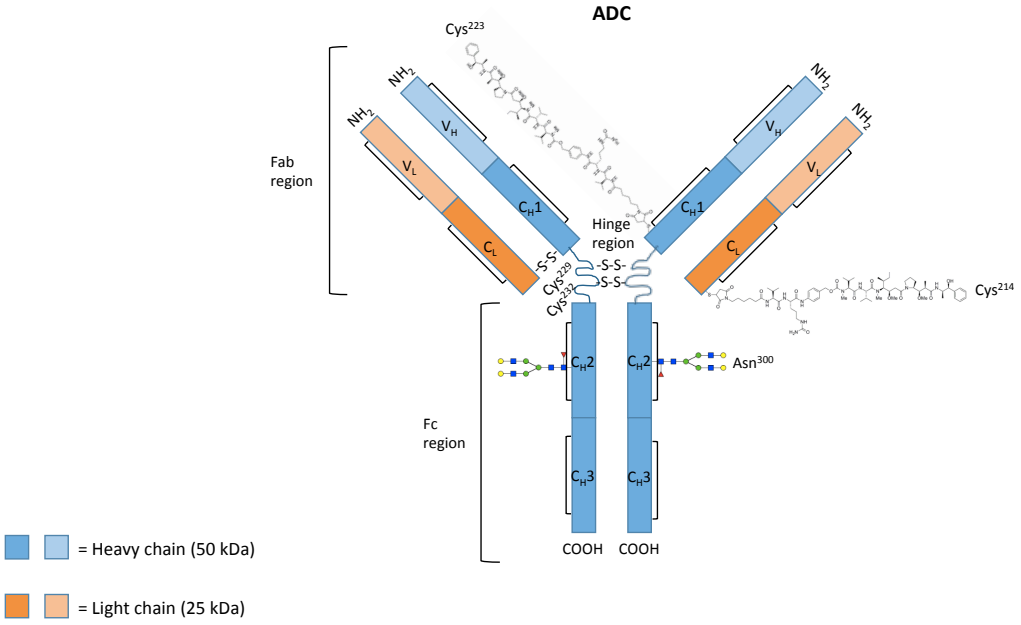


Figure 3.14: The antibody drug conjugates provided by ADC Biotechnology presented the above structure. The glycans are attached on the residue Asn (300) and the linker-drug molecules (vcMMAE) on Cys (214) and Cys (223) of the light and heavy chain respectively. Each antibody, therefore, can bring up to 4 drug molecules (here, only 2 are shown for clarity).

The identified **drug-peptides** were HC222–225, HC222–244, LC208–214 and LC212–214. While the peptides HC222–225 and LC212–214 were identified only in their drug-conjugated form (Appendix A, Figure A3, Figure A4), the peptides HC222–244 and LC208–214 appeared both as conjugated and non-conjugated. Figure 3.15 shows the spectrum of the non-conjugated peptide HC222–244 with m/z 844.04 for the ADC (S) sample, where all three cysteines are carboxymethylated (Cmc) and the b and y'' fragment ions are assigned. The respective spectra for all four samples seem quite similar qualitatively, and can be found in Appendix A (Figure A5). In Figure 3.16, the ADC spectra for the same peptide—but drug-conjugated—and the extracted

chromatograms for all four samples are shown. In this case two carboxymethylcysteines were identified, and one more cysteine (Cys (223)) that was linked to vcMMAE, resulting in m/z 1263.29. Only two y'' fragment ions were assigned, as the remaining energy from the electrospray ionisation is probably not sufficient to cause further fragmentation. The comparison of the four extracted ion chromatograms makes apparent the absence of such a peptide from the mAb samples. The non-conjugated peptide LC208–214 with m/z 870.34 and the assigned fragment ions are shown in Figure 3.17 for the ADC (S), while all four samples present a high degree of similarity as far as quality is concerned (Appendix A, Figure A6). The same peptide but with vcMMAE attached to Cys (214) (m/z 1064.56) is shown in Figure 3.18 for the ADC samples. As expected, the peak at ~23.4 min—where LC208–214 elutes—is absent from the extracted chromatograms of the mAb samples.

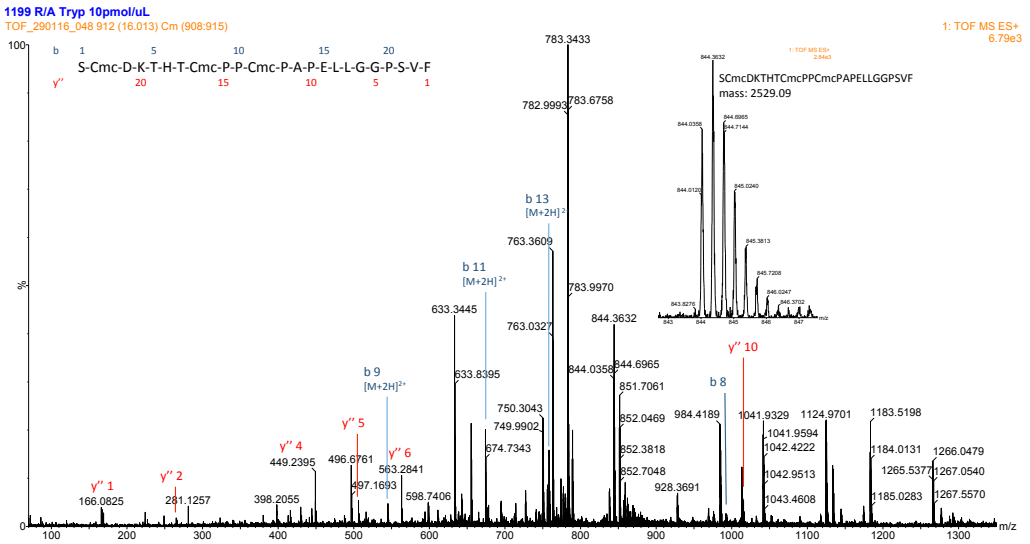


Figure 3.15: Spectrum of the non-conjugated peptide HC222–244 (RT 16.01 min) for the ADC (S) sample. All cysteines are carboxymethylated (Cmc) resulting in m/z 844.04 for the $[M+3H]^{3+}$ state. The observed b (blue) and y'' (red) fragment ions are assigned. A comparison of all four samples can be found in Appendix A (Figure A5).

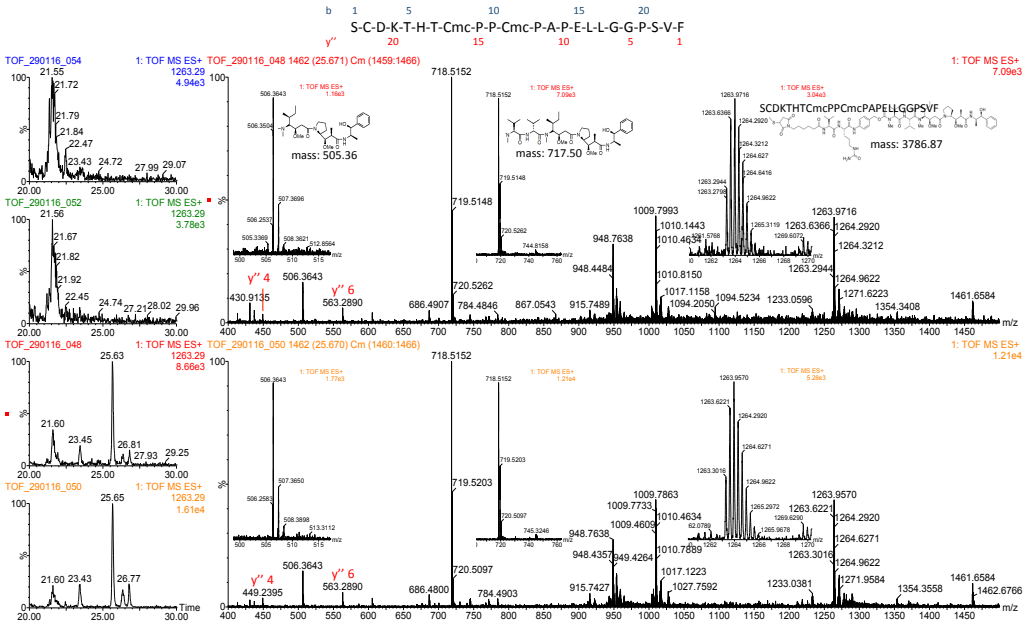


Figure 3.16: Spectrum of the drug-conjugated peptide HC222–244 (RT 25.67 min) for the ADC (S) (top, red) and ADC (LR) (bottom, orange) samples. One of the three cysteine residues (Cys (223)) is attached to the drug molecule resulting in m/z 1263.29 for the $[M+3H]^{3+}$ state. Signature fragment ions from vcMMAE are present as $[M+H]^{+}$, with m/z 718.51 (MMAE molecule) and 506.36. The extracted chromatograms on the left show the absence of the peak at ~25.60 min for the mAb samples (blue, green) in contrast to the ADC samples (red, orange). Only y'' fragment ions were observed and assigned to y''4 and y''6.

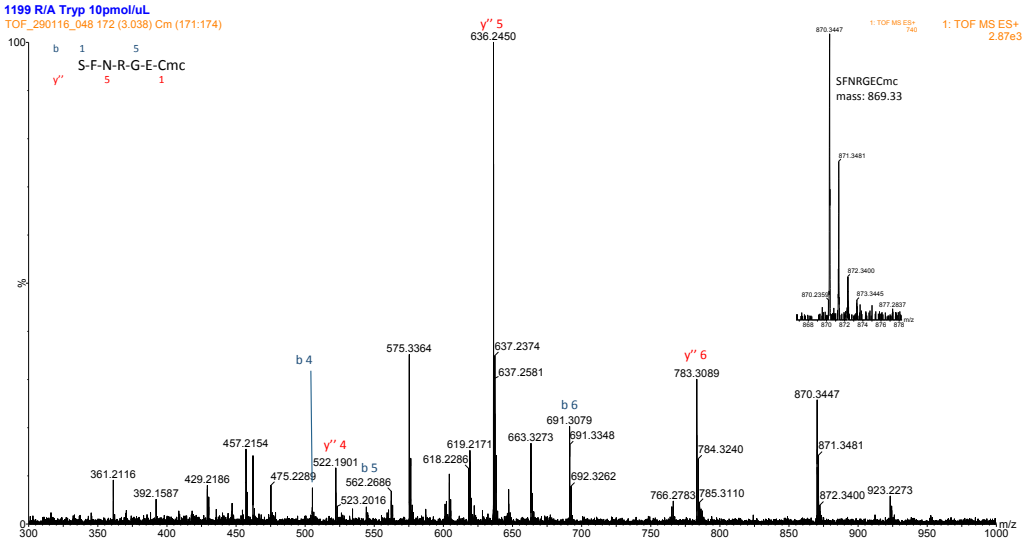


Figure 3.17: Spectrum of the non-conjugated peptide LC208–214 (RT 3.04 min) for the ADC (S) sample. Cys (214) is carboxymethylated (Cmc) resulting in m/z 870.34 for the $[M+H]^{+}$ state. The observed b (blue) and y'' (red) fragment ions are assigned. A comparison of all four samples can be found in Appendix A (Figure A6).

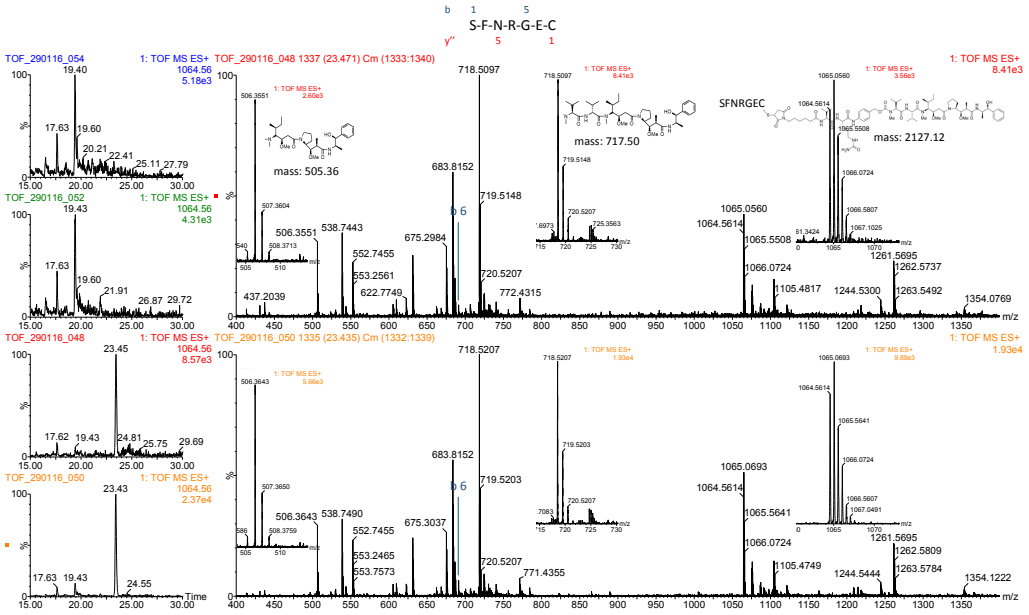


Figure 3.18: Spectrum of the drug-conjugated peptide LC208–214 (RT 23.45 min) for the ADC (S) (top, red) and ADC (LR) (bottom, orange) samples. The cysteine residue Cys (214) is attached to the drug molecule resulting in m/z 1064.56 for the $[M+2H]^{2+}$ state. Signature fragment ions from vcMMAE are present as $[M+H]^{+}$, with m/z 718.51 (MMAE molecule) and 506.36. The extracted chromatograms on the left show the absence of the peak at ~23.45 min for the mAb samples (blue, green) in contrast to the ADC samples (red, orange). Only the b6 fragment ion was observed.

In order to identify sites that are prone to modifications such as deamidation and oxidation, the respective residues were changed to the modified forms, and signals from the modified peptides were manually searched. For the **deamidation** sites, all asparagines (N) followed by a glycine residue were converted to aspartic acid (D) resulting in a mass increase of 1 Da, as explained above (Section 3.1.4), and all three of them were found on the heavy chain (N55, N318, N387). The identified peptides, found both in the native and modified form, were HC51–59, HC305–320, HC305–323 and HC374–395, with the native peptides eluting earlier, as asparagine is more hydrophilic than aspartic acid.[77] Figure 3.19 shows peptide HC51–59 and the observed fragment ions for ADC (S), while the spectra for all four samples and their extracted chromatograms can be found in Appendix A (Figure A7). As can be seen from the extracted chromatograms the native peptide eluted at ~5.9 min while the deamidated peptide appeared at ~6.1 min and ~6.4 min probably corresponding to the Asp and isoAsp forms. The weak intensities though do not allow for further confirmation through the fragment ions (Appendix A, Figure A8). The modified peptide HC305–320 was found at ~17.7 min with m/z 905.00 (Figure 3.20 and Appendix A, Figure A9). The glutamine residue Q314 is also a potential deamidation site. However, the presence of the fragment ions y''7 and y''8 are a proof that glutamine is intact in this case. The modified N318 was also identified in the peptide HC305–323 and can be

found in Appendix A (Figure A10 for the ADC (S) with the assigned fragment ions and Figure A11 for all four samples and their extracted chromatograms). The fragment ions y''10 or b10 were too large in this case to be identified and confirm for the native form of Q314. The modified form of HC374–395 was found at ~15.1 min with m/z 1273.06, showing the occurring deamidation on N387 (Figure 3.21 and Appendix A, Figure A12). For the same deamidated peptides, the succinimidyl intermediate was observed at a later RT than the aspartic acid form and with a mass decrease of 18 Da (Table 3.8 and Appendix A, Table A5–Table A8).

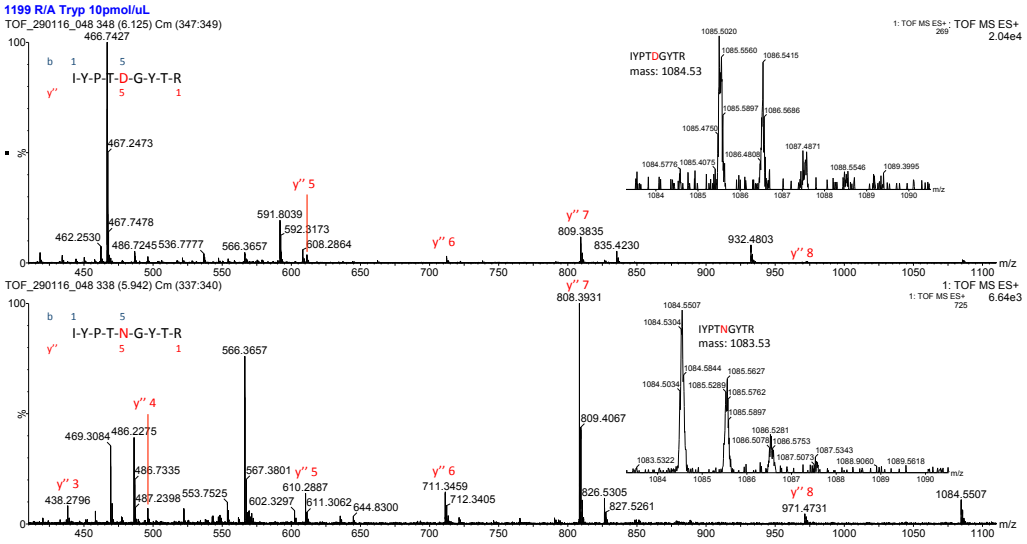


Figure 3.19: Spectra of the native (bottom) and deamidated (top) peptide HC51–59 with RT 5.94 min and 6.12 min respectively are shown for the ADC (S) sample. N55 is modified to aspartic acid resulting in a mass increase of 1 Da and m/z 1085.53 for the $[M+H]^+$ state. The observed y'' fragment ions are assigned. A comparison of all four samples can be found in Appendix A (Figure A7).

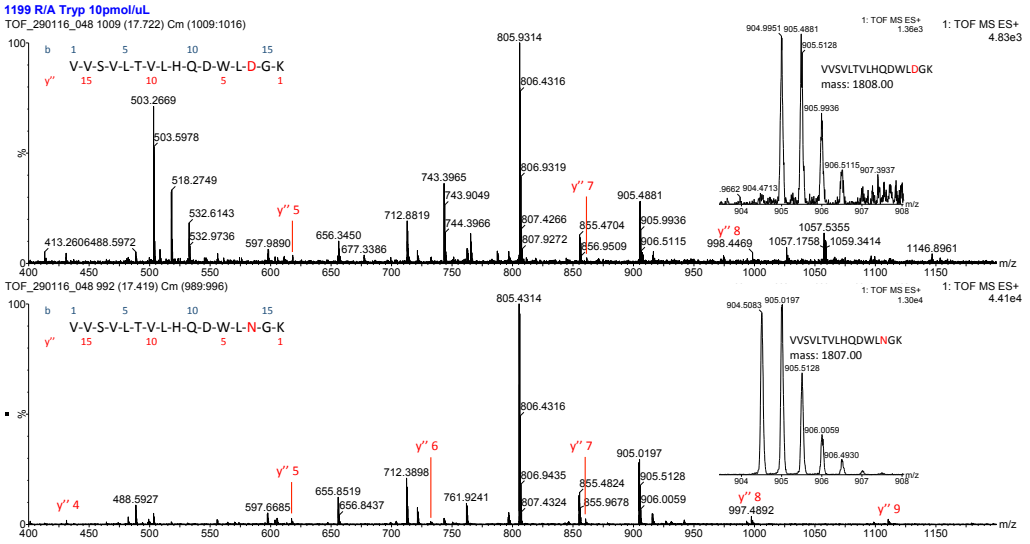


Figure 3.20: Spectra of the native (bottom) and deamidated (top) peptide HC305–320 with RT 17.42 min and 17.72 min respectively are shown for the ADC (S) sample. N318 is modified to aspartic acid resulting in a mass increase of 1 Da and m/z 905.00 for the $[M+2H]^{2+}$ state. The observed y'' fragment ions are assigned. A comparison of all four samples can be found in Appendix A (Figure A9).

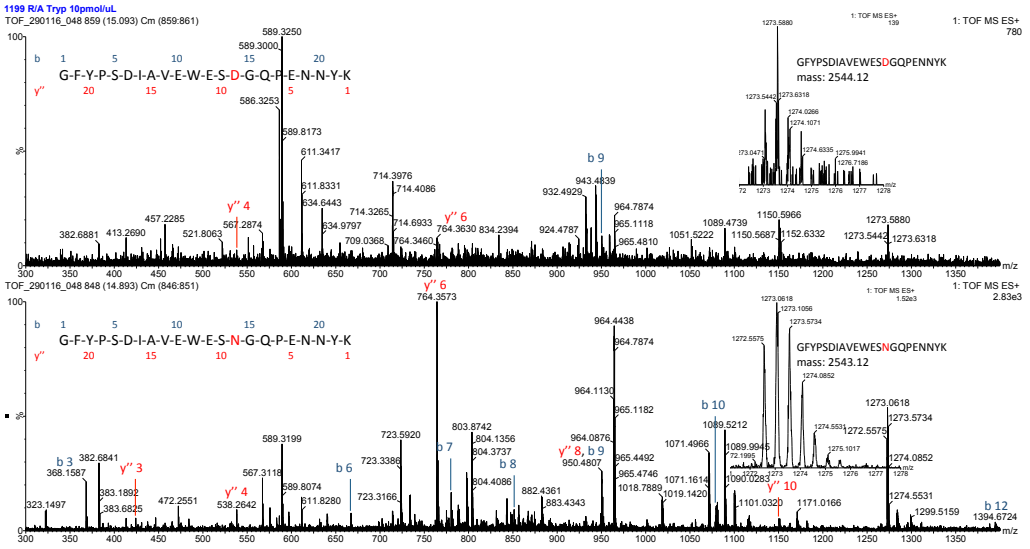


Figure 3.21: Spectra of the native (bottom) and deamidated (top) peptide HC374–395 with RT 14.89 min and 15.09 min respectively are shown for the ADC (S) sample. N387 is modified to aspartic acid resulting in a mass increase of 1 Da and m/z 1273.06 for the $[M+2H]^{2+}$ state. The observed b (blue) and y'' (red) fragment ions are assigned. A comparison of all four samples can be found in Appendix A (Figure A12).

For the identification of the **oxidation** sites, all methionines (M) were converted to methionine sulfoxide (M*) in the used software, resulting in a mass increase of 16 Da, and one of them was found on each chain (M4 on the light chain and M255 on the heavy chain). The identified peptides—present both in the native and modified form—were LC1–18 and HC252–258, with the native peptides eluting later, due to the higher hydrophobicity of methionine compared to methionine sulfoxide.[78]

Figure 3.22 shows peptide LC1–18 and the observed fragment ions for the ADC (S), while the spectra for all four samples and their extracted chromatograms can be found in Appendix A (Figure A13). The elution time for the modified peptide was ~8.1 min with m/z 947.93, while the native peptide appeared ~2.6 min later (RT 10.7 min). The fragment ion b4 in the spectrum of the oxidised peptide confirms the presence of the modified M*4. Peptide HC252–258 eluted at ~5.9 min with m/z 851.45 (Figure 3.23 and Appendix A, Figure A14) showing the occurring modification on M255, while the native form had an elution time of ~6.9 min. The presence of the fragment ions b4, y''4 and y''5 in the spectrum of the oxidised peptide further confirms the modification.

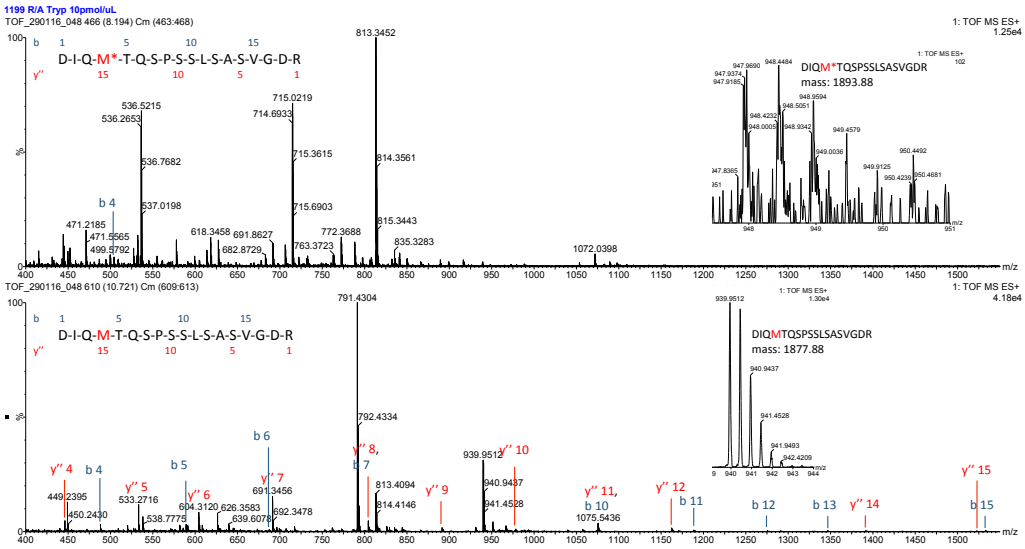


Figure 3.22: Spectra of the native (bottom) and oxidised (top) peptide LC1–18 with RT 10.72 min and 8.19 min respectively are shown for the ADC (S) sample. M4 is modified to methionine sulfoxide resulting in a mass increase of 16 Da and m/z 947.93 for the $[M+2H]^{2+}$ state. The observed b (blue) and y'' (red) fragment ions are assigned. The fragment ion b4 in the oxidised spectrum confirms the modification. A comparison of all four samples can be found in Appendix A (Figure A13).

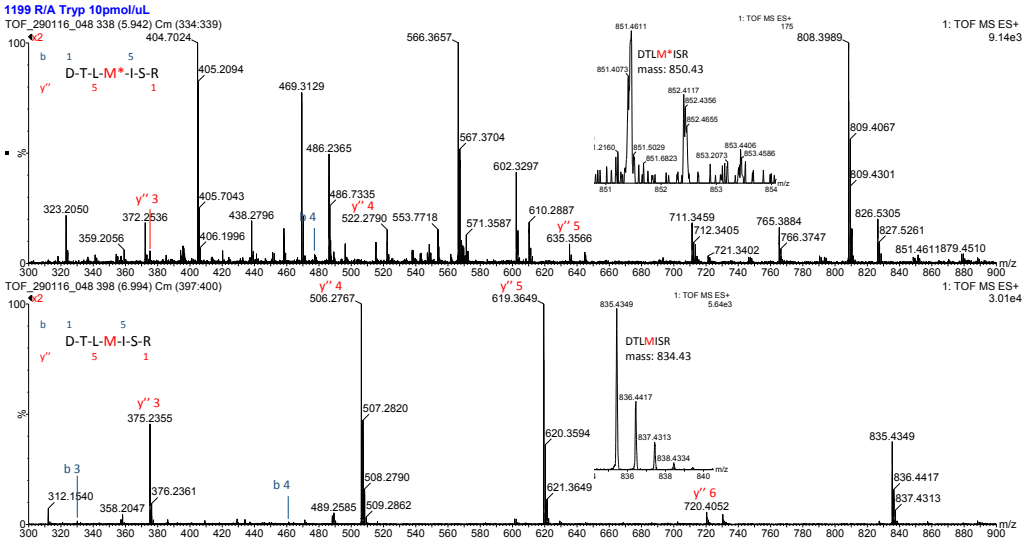


Figure 3.23: Spectra of the native (bottom) and oxidised (top) peptide HC252–258 with RT 6.99 min and 5.94 min respectively are shown for the ADC (S) sample. M255 is modified to methionine sulfoxide resulting in a mass increase of 16 Da and m/z 851.45 for the $[M+H]^+$ state. The observed b (blue) and y'' (red) fragment ions are assigned. The fragment ions b4, y''4 and y''5 in the oxidised spectrum confirm the modification. A comparison of all four samples can be found in Appendix A (Figure A14).

The relative amounts of oxidation, deamidation and the succinimidyl intermediate were calculated for each sample and are presented in Table 3.8, while a more detailed description of the integrated areas of the chromatogram peaks for the modified and native peptides is provided in Appendix A (Table A5–Table A8). The amount of oxidation on M4 was calculated as ~1%, while M255 appeared to be more susceptible (~3%).[55,74] Deamidation on N55 and the succinimidyl intermediate varied among the four samples from 17–27% and 5–9% respectively. For N318 the average deamidation resulting from the two peptides was calculated as ~13% and the presence of the succinimidyl intermediate was estimated as 2%. Only the mAb (LR) presented a higher amount of deamidation on N318 (17% on average), which resulted from peptide HC305–323. Finally deamidation on N387 varied from 7% (for mAb (S)) to ~12% (for the rest of the samples) and the succinimidyl intermediate was only identified in the mAb (S) as 9%.

Table 3.8: The amount of oxidation, deamidation and the succinimidyl intermediate is presented as a percentage (%) of the total area (sum of the modified and native form).

Site	Peptide	Comment	mAb (S)	mAb (LR)	ADC (S)	ADC (LR)
M4	LC1–18	Oxidation	0.7	0.4	1.0	0.4
M255	HC252–258	Oxidation	1.9	3.3	3.4	2.0
N55	HC51–59	Deamidation	16.9	20.3	26.9	19.2
		Succinimidyl	4.9	5.8	8.6	6.4
N318	HC305–320	Deamidation	5.8	6.9	6.4	5.9
		Succinimidyl	1.1	1.2	1.1	0.9
	HC305–323	Deamidation	17.6	29.2	18.9	18.6
		Succinimidyl	2.6	2.7	2.9	2.6
N387	HC374–395	Deamidation	6.7	11.7	11.4	12.3
		Succinimidyl	8.9	N/A	N/A	N/A

3.3.2. TQ-S results

As described above (Q-TOF results, Section 3.3.1), for peptide mass analyses, the protein should be subjected to a treatment that would render its structure more accessible to enzyme digestion. In this case, the ADC samples were treated with CNBr, which cleaves at the C-terminal end of methionine (M) while the disulfide bonds remain intact. The reaction was performed in darkness, in order to prevent modification of the methionines, as exposure to aqueous environment increases the possibility of oxidation. Tryptic digestion followed and the resulting peptide mixture was separated by reversed-phase UPLC. The selected peptides were directed into the tandem quadrupole mass spectrometer (MS/MS) and the transitions to their fragment ions were monitored in order to facilitate the quantitative analysis. The TIC plots from the eight repetitions for each sample (Appendix A, Figure A15, Figure A16) showed that LC46–61 (RT ~22.3 min) was the most abundant and reproducible peptide and was therefore chosen as the marker peptide. The drug-conjugated peptides HC222–225 (RT ~30.3 min) and LC212–214 (RT ~32.7 min) were monitored in two transitions. The integrated areas for each one of the monitored transitions are given in Appendix A (Table A9) and the relative abundance of the drug-conjugated peptides for each sample is shown in Figure 3.24. Conjugation on the light chain seems to be comparable for the two samples with a ratio to the marker peptide of ~2.5 for the ADC (S) and ~2.2 for the ADC (LR) on average. The heavy chain for the ADC (S) presents approximately 2.5 times the conjugation of the ADC (LR) and the calculated ratios are ~1.0 and ~0.4 respectively. For both chains, conjugation on the ADC (LR) presents better precision perhaps indicating that the lock-release method is more controlled. The light chain

responds better to MS, therefore resulting in a significant difference between the two chains and preventing direct comparison.

Data acquired by ADC Biotechnology showed a DAR (drug-to-antibody ratio) of 3.6 for the ADC (S) and 3.7 for the ADC (LR). In addition, the lock-release method appeared to be of higher purity resulting in <0.5% of free drug, while in the solution method 4.5% of free drug was present. The amount of free drug in antibody-drug conjugate preparations is important to be maintained at low levels, as due to its high toxicity it can potentially affect the safety of the final product.

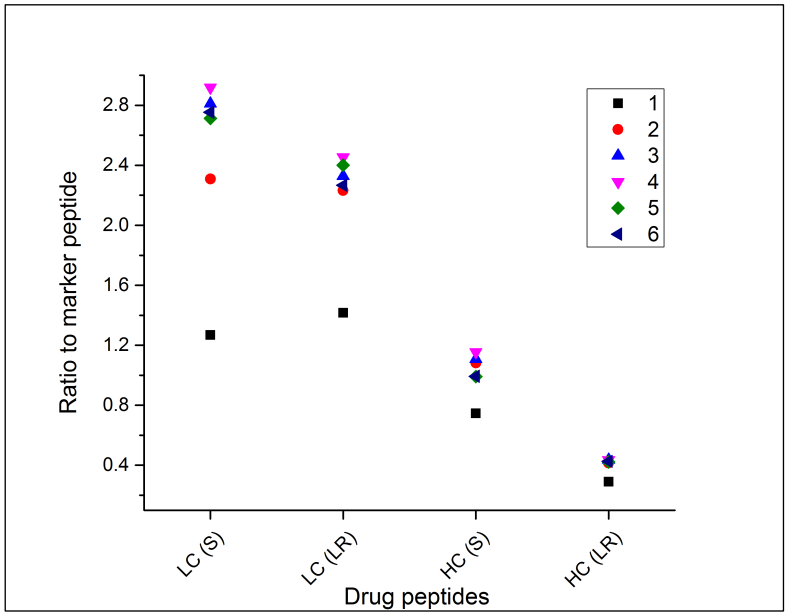


Figure 3.24: Relative abundance of the drug-conjugated peptides, calculated by dividing the sum of the areas from the two transitions by the marker peptide area. LC and HC are the light and heavy chains respectively. (S) and (LR) are the ADC samples produced with the solution and the lock-release method. The 6 presented measurements were acquired from three replicates for each sample (ADC (S) and ADC (LR)).

3.4. DLS SPECTROSCOPY

Previous studies on the stability of trastuzumab, using DLS spectroscopy among other techniques, have shown a hydrodynamic diameter of 11.2–11.6 nm for the native antibody.[79,80] In this study, the hydrodynamic diameter was measured as 10.8–11.7 nm (intensity values), with the mAb (S) presenting the lowest value and the ADC (S) the highest, while mAb (LR) and ADC (LR) were both measured as 11.5 nm (Figure 3.25, left). This variation in the measured values between the mAb (S) and the rest of the samples is probably the result of higher inherent stability and/or purity of the first. The right image in Figure 3.25 shows the size distributions at the end of the temperature experiment, after the samples were heated up to 100 °C and subsequently cooled down to 20°C. As can be easily observed, mAb (S) was the only sample that has retained particles of the initially detected size, indicating higher structural stability than the rest of the samples, and especially than the ADC (S), which presented particles of larger size and lower intensity, indicative of its decreased stability. All four samples presented wider peaks when cooled down to 20°C, suggesting the appearance of larger particles, which affect the shape of the distribution. As described in Section 1.2, and also discussed below in the CD Section (3.5.1), application of high temperatures leads to activation of the energy barriers between the native and the denatured states. However, only partial unfolding of the antibodies took place here, since the DLS temperature-dependent experiment was conducted up to 75 °C, namely below the temperature that the completely unfolded structure was observed in the CD results (83 °C). As explained in Section 3.5, the different domains of trastuzumab unfold in two stages (at two different temperatures). It seems, therefore, that only the first stage was satisfied during the DLS experiment (especially for the conjugated samples that presented an earlier conformational change), and the partially unfolded structure ended up to a misfolded intermediate (expanded structure) in one of the local minima in the energy funnel after the samples were cooled down. In the case of the conjugated samples, intermolecular interactions (aggregation) might have also taken place, however to a limited extent, since the size of the particles for the two ADCs at the end of the experiment was relatively small (close to the initial value) and therefore not sufficient to suggest with certainty that aggregation occurred. The intensity distributions of the samples at each temperature can be found in Appendix A (Figure A17–Figure A20).

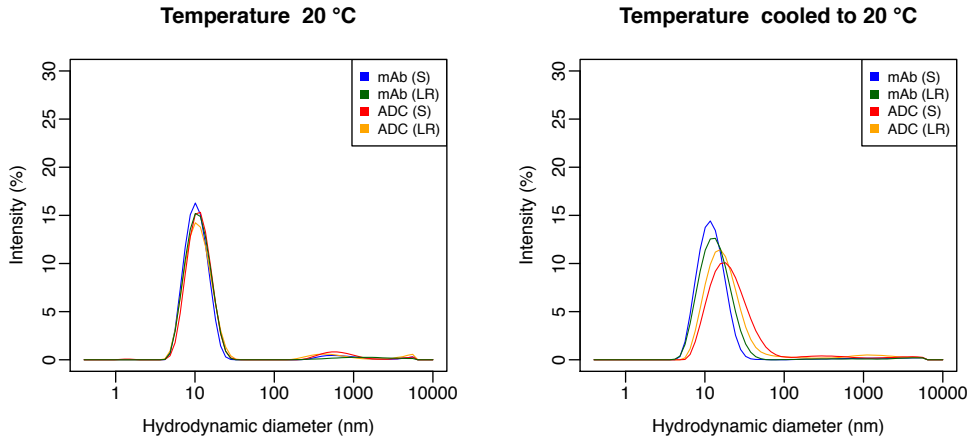


Figure 3.25: Size distributions presented in intensity values (nm) for the four samples at 20°C (left), and after the samples were cooled down to 20°C (right). The distributions become wider at the end of the experiment and the ADC samples shift to larger values. Both observations indicate the appearance of larger particles. The presented data are averaged over 6 replicates.

Intensity values are highly affected by the presence of larger particles and impurities, shifting the distribution to higher values. Since the samples presented moderate uniformity in the majority of the measurements, with polydispersity indices (PDI) between 0.1 and 0.4, the results are henceforth presented in number values as more representative of the most abundant species in the sample. Figure 3.26 (left) shows the mean size of the particles observed in each sample (averaged over six replicates) and how this is changing over the temperature range 20–75°C and when the sample is cooled down to 20°C. The mAb samples seem to follow a common path throughout the experiment, showing a slight increase in size from ~6.8 nm at 20 °C to 7.1 nm when cooled down to 20°C. The ADC samples also present very similar trends to each other, but different from that of their native relatives. The size of the ADC (LR) was measured as 6.6 nm at 20°C, while ADC (S) presented the largest value of all four with a mean particle size of 7.2 nm. Both conjugated samples appear to be more prone to the appearance of larger particles with sizes between 8.5–8.9 nm at the end of the experiment (cooled to 20°C), perhaps indicating expansion of the structure rather than prominent aggregation, as aggregation would be expected to lead to much larger values.

The standard deviations with a 95% confidence interval were calculated over the averaged number values of the six replicates and are shown as error bars in Figure 3.26 (right). The sample mAb (LR) presented a large deviation from 30–45°C, due to a faulty replicate that presented particles up to 10.8 nm. This effect is also reflected in the mean particle sizes at the specific temperatures. The rest of the samples also

presented larger particles than the given mean sizes, occasionally, in the order of 9.4–10.2 nm. Since this was only observed in one of the replicates in each case, the effect is considered insignificant and the mean particle sizes are a good approximation. However, the effect is reflected in the calculated standard deviations. The mean particle size values with their standard deviations at each temperature can be found in Appendix A (Table A10). Finally, very small particles (~3.0 nm) appeared in one of the replicates of mAb (LR) and two of the replicates of ADC (S) at temperatures above 40°C. Again these observations were not consistent—as the particles appeared at one or two temperatures only—and therefore they were not included in the analysis.

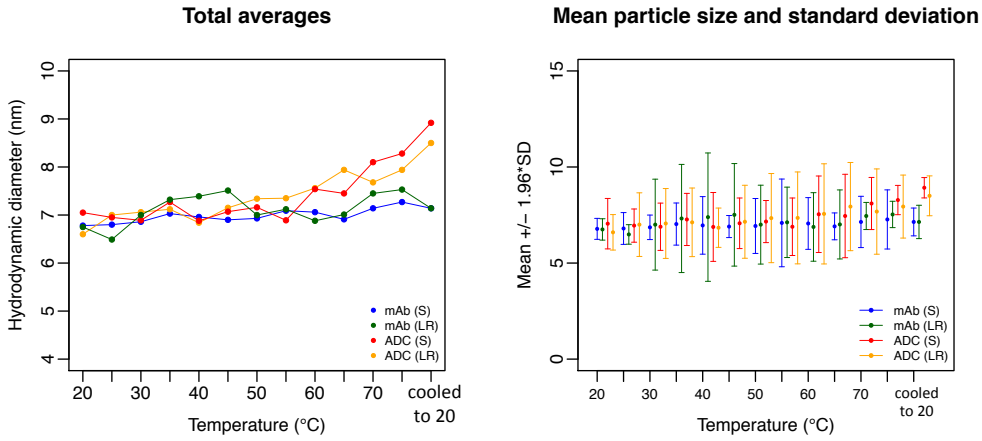


Figure 3.26: Hydrodynamic diameter (nm) of the four trastuzumab samples over a range of temperatures from 20–75°C and after the samples were cooled down to 20°C. Left: the averaged sizes over 6 replicates are shown. Right: overall mean particle diameters (solid circles) and standard deviation with a 95% confidence interval (error bars). The results are presented in terms of number values.

In conclusion, the DLS data showed that native trastuzumab molecules (mAb samples) have an inherently high structural stability at temperatures up to 75°C, presenting only a small increase in their size. The ADC samples were less stable than the native antibodies, showing a significant and irreversible expansion of their structure at temperatures from 60°C onwards. This structural instability is probably acquired during the stage of reduction of the inter-chain disulfide bonds, which precedes the conjugation process. However, the whole structure is still held together by noncovalent bonds, which prevent the structure from collapsing into smaller molecules (*e.g.* fragments composed of different combinations of the heavy and light chains as implied by Figure 3.4). Furthermore, the ADC (S) sample presented a wider distribution in comparison to that of the ADC (LR) perhaps indicating the presence of

impurities that affect the measured intensity values. As a result, lock-release seems to have an advantage over the solution method, providing a final product of higher purity. Finally, neither of the samples presented any prominent aggregation, which could potentially be explained by the protective effect of the formulation excipients.

3.5. CD Spectroscopy and Secondary Structure Estimation

Previously published studies on the thermal stability of IgG antibodies and more specifically on trastuzumab have used CD spectroscopy[79] and differential scanning calorimetry (DSC).[81,82] These studies have shown that IgG antibodies present two (or more) transition states of unfolding, which correspond to unfolding of different domains or groups of domains. The melting temperatures however vary significantly between murine and humanised antibodies, among different isotypes—as flexibility of the hinge region differs—and even among changing variable domains of the same isotype. Trastuzumab has appeared to unfold in two stages, with the first occurring at ~70 °C due to the C_H2 domain and the second at ~82 °C due to unfolding of the Fab and C_H3 domain of the Fc fragment.[82]

3.5.1. CD spectra of trastuzumab

In the present study, spectra were obtained for the trastuzumab samples at their original **concentrations** (~1 mg/mL) using a ~0.01 cm path length demountable cuvette (Figure 3.27, left), and following 10-fold dilution using a 0.1 cm path length fixed cuvette (Figure 3.27, right). The data were obtained at 20 °C and are the average of 5 and 6 replicates respectively. The normalised data look almost identical at the two different concentrations, presenting the typical β-sheet spectrum with a positive peak at ~201 nm and a negative peak at ~217 nm, characteristic of the antibody. The "shoulder" at ~232 nm is probably a contribution from the turns spectrum (usually negative at ~190 nm, positive at ~208 nm and a low negative signal at ~228 nm). The distortion of the spectra below 200 nm for the 1 mg/mL samples is due to the distorted HT signal (flattened, data not shown) at lower wavelengths. At both concentrations, mAb (S) presented a slightly stronger signal at 217 nm, while ADC (S) was somewhat weaker than the rest of the samples in the same area, which could probably be a first sign of the difference in inherent stability among the samples. A more detailed comparison between the two concentrations for each sample is provided in Appendix A (Figure A21). The high concentration spectra were recorded on a J-815 spectrophotometer, while the 0.1 mg/mL data were recorded on a J-1500 model. In order to correct for the variance in the calibration between the two instruments, the 1 mg/mL spectra were shifted by −3 nm. The similarity of the spectra at the two concentrations confirms that there are no significant differences in the structure of any of the four samples after dilution. Therefore, the following temperature experiments were performed with the

diluted forms, due to the convenience of a fixed path length cuvette. Table A11 (Appendix A) shows the maximum, minimum and zero-crossing points for the 6 replicates of the 0.1 mg/mL data at 20 °C.

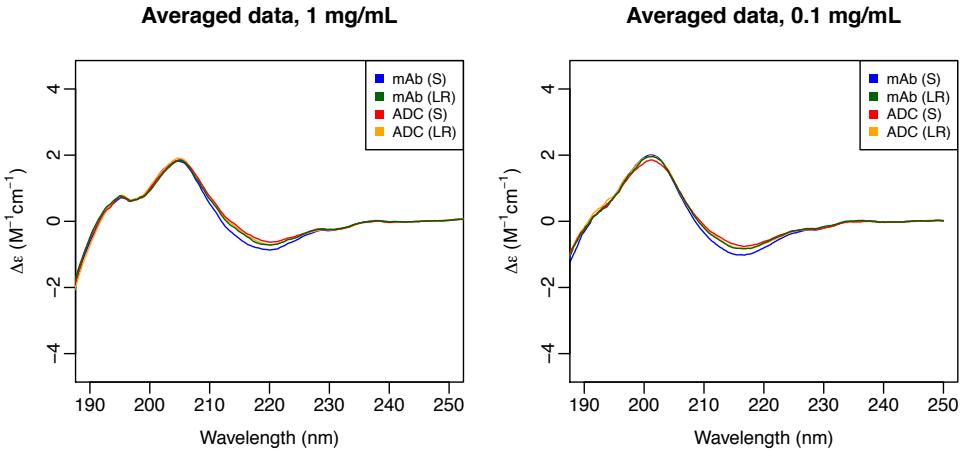


Figure 3.27: CD spectra of the four trastuzumab samples at their original concentration (~1 mg/mL) using a 0.01 cm demountable cuvette (left) and after 10-fold dilution (0.1 mg/mL) using a 0.1 cm fixed cuvette (right). All data were acquired at 20 °C. The high concentration spectra are the average of 5 replicates (16 scans each), and the diluted spectra are averaged over 6 replicates (16 scans each).

The used **formulation** has a relatively strong CD signal, which originates from the His-Cl buffer in the wavelength region from ~198 nm onwards, and from trehalose at lower wavelengths (Figure 3.28, left). The formulation spectrum was also checked for any changes at higher temperatures (up to 100 °C) and a gradually decreasing intensity was observed, while the shape of the spectrum remained exactly the same (Appendix A, Figure A22, left). At the end of the temperature experiment, and after the cuvette was cooled down to 20 °C, the formulation spectrum was identical with the initial one at 20 °C (Appendix A, Figure A22, right). For the baseline correction of the trastuzumab spectra at different temperatures, the blank (formulation only) spectra at 20 °C were used (Figure 3.28, right). Recording the blank in replicates showed a variance, which is probably the result of a combination of reasons such as different cuvettes, noise from the instrument and different days of acquiring the data. For the first three replicates the same blank spectrum (Bl1) was used in all cases, while for the last three replicates a blank spectrum recorded on the same cuvette and on the same day as its respective sample was used (Bl2* replicates) in order to correct for the

* Replicates with 4 scans were used for the spectra at different temperatures and replicates with 16 scans were used for the spectra at 20 °C, according to the spectra of the samples.

variation. However, the total averages of the Bl2 replicates seem to eliminate the observation and the Bl1 and Bl2 (total average) spectra seem to overlay sufficiently in the region from 200-220 nm, where the samples differ the most.

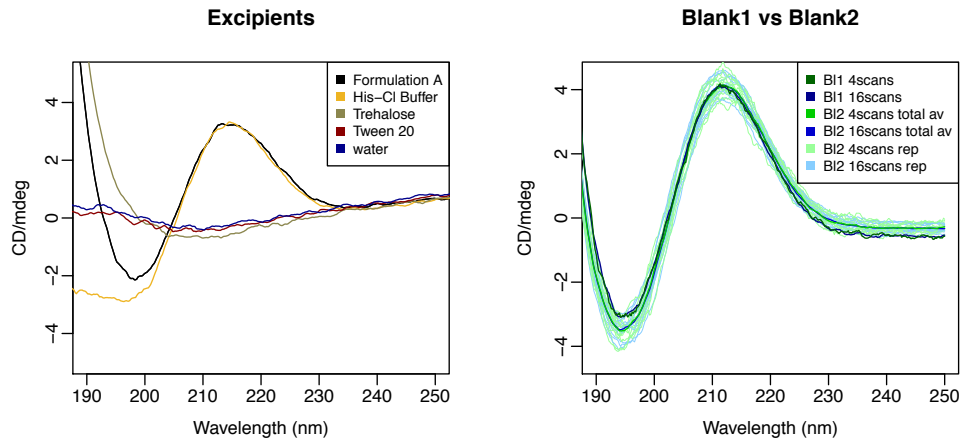


Figure 3.28: Left: CD spectra of the formulation and for each one of the excipients at their original concentrations (5 mM His-Cl buffer pH 5.95, 50 mM trehalose, 0.01% w/v Tween 20). The spectra were recorded on a J-815 spectrophotometer, using a demountable cuvette with a nominal path length of 0.01 cm. Right: CD spectra of the formulation replicates after 10-fold dilution. The spectra were recorded on a J-1500 spectrophotometer, using a 0.1 cm fixed path length cuvette according to the following temperature experiment. Bl1 (=Blank 1) was only recorded once (with 4 and 16 scans) and was used for correcting the spectra of the first three replicates. Bl2 (=Blank 2) spectra were recorded in replicates, on the same day and cuvette as the matching sample spectrum, and they were used for correction of the last three replicates. The total averages of the spectra for the Bl2 replicates are also given. The 16 and 4 scans spectra were used for correction of the spectra at 20 °C and at different temperatures respectively, in accordance with the spectra recorded for the samples.

The averaged CD spectra of the samples at increasing **temperature** from 20 °C to 100 °C, per 1 °C, are shown in Figure 3.29. As the temperature increases, the characteristic β-sheet spectrum of the antibody is gradually changing towards a spectrum with a negative peak at ~204 nm, indicating that the structure is dominated by random coil character (usually negative at ~198 nm and a weak positive signal at ~212 nm). All four samples seem to be quite stable up to a temperature of 54 °C, when the conjugated antibodies start to differentiate from the native antibodies (see below, Figure 3.30), while the native antibodies start to diverge from their initial spectrum at ~65 °C (see below, melting curve Figure 3.31). The spectrum that corresponds to the first unfolded structure in each case is emphasised by a bold line and occurs at different temperatures for the native and the conjugated antibodies, revealing a difference in their inherent stability. Both mAb (S) and mAb (LR) seem to reach their unfolded structure at 82 °C (grey), while the ADC samples appear to be slightly less stable with an unfolding temperature at 80 °C (purple). From the unfolding temperature onwards,

the spectra are hardly changing. A table with the specific unfolding temperatures for each replicate is given in Appendix A (Table A12).

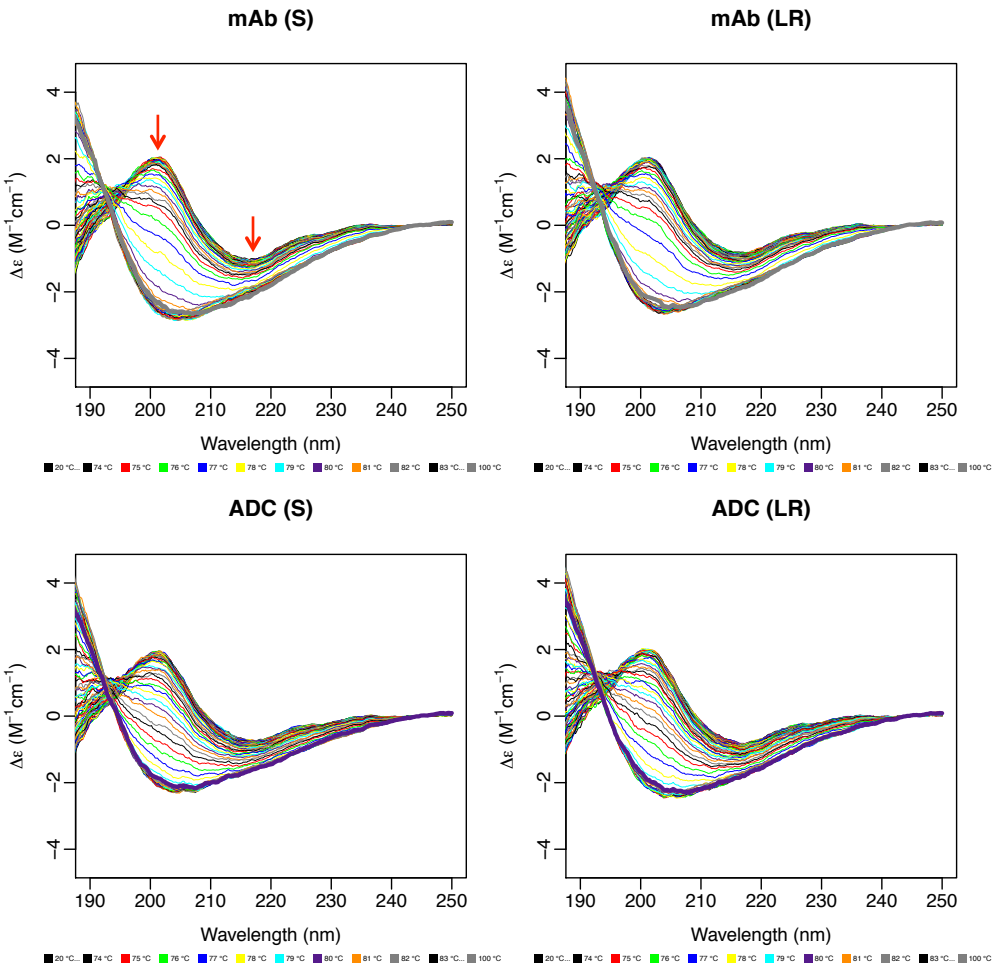


Figure 3.29: CD spectra of the trastuzumab samples after 10-fold dilution at a temperature range of 20–100 °C, per 1 °C. The spectra were recorded on a J-1500 spectrophotometer, using a 0.1 cm fixed path length cuvette. The red arrows in the plot of the mAb (S) sample show the direction of the decreasing and increasing signal intensities at 201 nm and 217 nm respectively with increasing temperature, which was observed for all four samples. The first unfolded spectrum for each sample is emphasised, showing the different temperatures that unfolding takes place between the native (at 82 °C, grey) and the conjugated (at 80 °C, purple) antibodies. The spectra are averaged over 6 replicates.

Figure 3.30 provides a comparison of how **unfolding** is progressing for the four samples at selected temperatures. As mentioned above, the CD spectra of the two conjugated antibodies start to differentiate from the native antibodies at ~54 °C, with the ADC samples gradually losing intensity at 201 nm. After the remaining shoulder at 201 nm is completely lost (at ~72 °C for the ADC and ~78 °C for the mAb samples), the negative peak increases in intensity and at the same time shifts from 217 nm to 204 nm. The spectra remain different until 83 °C, when all samples are unfolded and

appear very similar in shape, but different in intensity. As with the 20 °C spectra, mAb (S) presents a stronger signal while ADC (S) has the weakest signal of all four. From 83 °C to 100 °C there is hardly any change in the CD spectra of the trastuzumab samples, perhaps indicating that the completely unfolded state is reached after activation of the energy barriers between the native and the denatured state. Some variability was observed in one of the mAb (S) replicates in the temperature range 76–80 °C (data not shown), though not significant enough to affect the final conclusion.

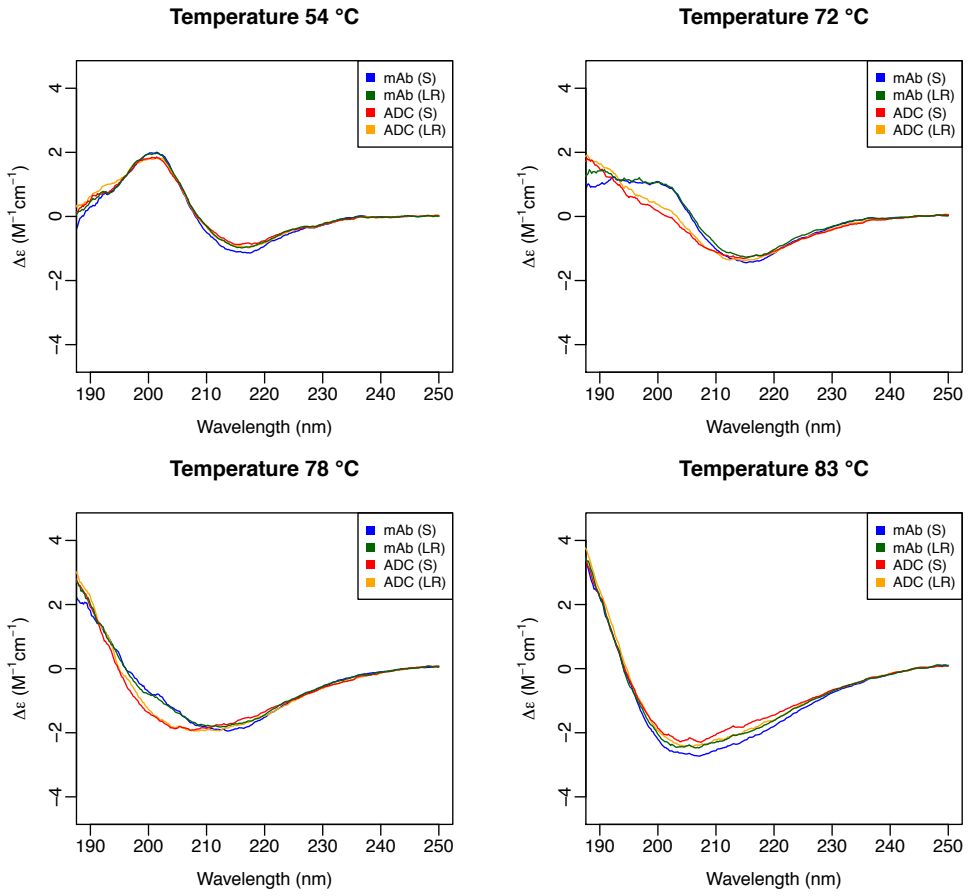


Figure 3.30: Comparison of the CD spectra of the trastuzumab samples at selected temperatures, showing the differentiation of the conjugated from the native antibodies. All four samples are unfolded at 83 °C, therefore presenting very similar spectra but different intensities. The data are averaged over 6 replicates.

The signal at 201 nm was followed throughout the temperature experiment for all four samples and the extracted **melting curves** are shown in Figure 3.31 (left) as an average of 6 replicates. As discussed above and becoming more obvious from the melting curves, the ADC samples start changing at ~54 °C, while the mAb samples seem to diverge from their initial spectrum at ~64 °C. The conjugated samples obtain

their unfolded structure at ~82 °C, while the native antibodies follow later, at ~83 °C. At the end of the experiment, and after the samples were cooled down to 20 °C (right), the negative peak appears much wider and shifted at ~210 nm perhaps denoting a combination of random coil and β-sheet structure or even aggregation. As described in Section 1.2, thermal denaturation leads to a high-energy unfolded state of the protein due to loss of the native interactions. Cooling of the samples (restoring the folding conditions) results in lowering the energy state to local minima, by folding of the protein towards misfolded intermediates or by utilising intermolecular interactions (aggregation). Both of these events are irreversible, especially in the case of large multidomain proteins such as monoclonal antibodies. The respective HT spectra, at 20 °C and after the samples were cooled down to 20 °C, are shown in Figure A23. The higher HT values of the samples after they returned to the initial temperature (cooled to 20 °C) denote the presence of scattering particles and further support the observation that aggregation occurred upon cooling of the samples. However, no solid particles were visible to the naked eye at inspection of the samples at the end of the experiment.

Though the two transition states are not that obvious in the present study, the point where 201 nm crosses zero (at 72 °C for the ADCs and 78 °C for the mAbs) could be regarded as the first state, and the completely unfolded spectrum (80 °C for the ADCs and 82 °C for the mAbs) as the second. Therefore in the case of the ADCs, the first state, at which C_H2 is supposed to unfold according to the literature, is probably affected by the increased flexibility of the hinge region due to the reduced inter-disulfide bonds, while the second transition state, at which Fab and C_H3 domains unfold, is probably affected by the presence of the toxin on the Fab region. In both cases the conjugated samples present reduced stability compared to the mAbs, and earlier unfolding of their distinct domains.

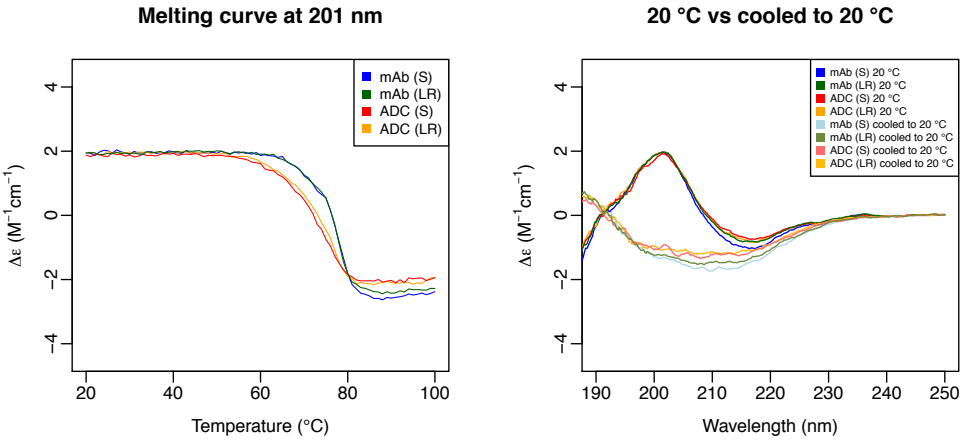


Figure 3.31: Left: Melting curves of the trastuzumab samples at 201 nm. The curves were extracted from the normalised data (Δε units) and averaged over 6 replicates. Right: Comparison of the CD spectra of the four samples at 20 °C and at the end of the experiment, after the samples were cooled down to 20 °C. Only the last 3 replicates were used for calculating the average.

3.5.2. Secondary structure estimation

The averaged spectra of the four samples were analysed for their secondary structure content at each temperature using SSNN. Figure 3.32 shows the estimated changes that occur in each sample, focusing on the main four structure types (α-helix, β-sheet, turns and other*). For all four samples the initial and the final estimations are almost identical (at 20–57 °C and 82–100 °C respectively), while from 58–81 °C some variability is observed—especially in β-sheet and other—between the conjugated and the native antibodies. The overall change in the antibodies' structure appears to be in the order of 20%, originating from a decrease in the β-sheet content (from 44% to 24%), and shared as an increase between other (30% to 37%) and α-helix (6% to 19%), while turns remain almost unaffected (20%). The increase in the α-helix content is probably an artefact which is reflected in the relatively small increase of "other", due to a misfit of the predicted to the original spectra from ~80 °C onwards. A false feature at ~228 nm is incorrectly estimated for all four samples (Appendix A, Figure A24). The SSNN predicted results for the spectra at 20 °C were compared to the crystal structure of an IgG2a antibody (PDB ID: 1IGT),[30] for which the structures were assigned as: α-helix 6%, β-sheet 45%, turns 9% and other 40%. Conclusively, the results for α-helix and β-sheet seem to be in good agreement with the crystal structure, while turns and other present a 10% difference.

* Other can refer either to unordered structure or to unassigned secondary structure.

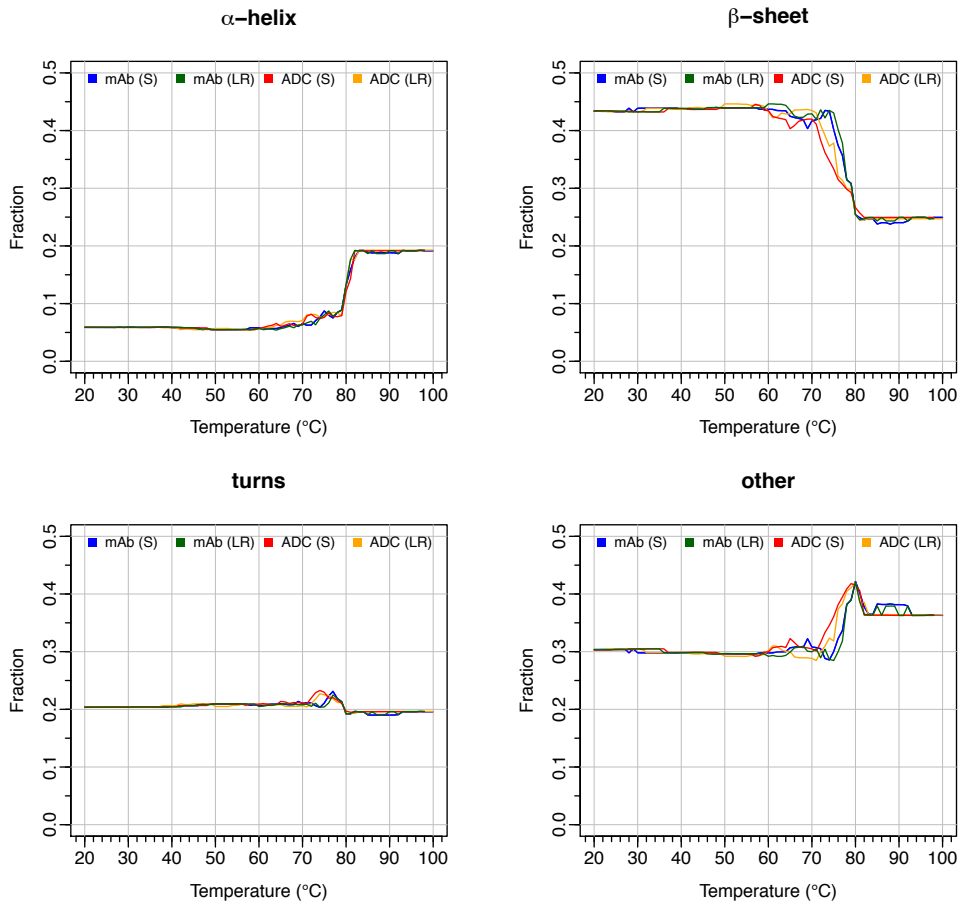


Figure 3.32: Results from the secondary structure analysis of the averaged trastuzumab spectra at a temperature range from 20–100 °C, per 1 °C. SSNN was used for the analysis, and the assigned structures are α-helix, β-sheet, turns and other, presented in fraction values.

3.6. CONCLUSION

MS, CD and DLS are three of the most useful techniques for physicochemical and structural characterisation of biotherapeutic molecules. Due to the distinct information that each one of them produces regarding the structure of the protein molecule, combination of all three is necessary in order to carry out a more complete study, when a biosimilar-to-innovator or batch-to-batch comparison is required. CD and DLS were coupled with the use of temperature dependence in order to aid interpretation and further confirm observations that were made at room temperature. Comparison of four trastuzumab samples—two in the native state and two conjugated with a toxic agent using different conjugation methods—showed differences in the inherent stability and some variation in purity.

MS showed some variability in homogeneity and purity in certain regions of the TIC plot of the four samples. However, glycosylation appeared homogeneous

among the samples and the relative amounts of modifications such as deamidation and oxidation presented comparable values, with the exceptions of ADC (S) and mAb (LR) that exhibited higher deamidation on N55 and N318 respectively. For both ADCs, drug conjugation took place on the heavy-light chain inter-disulfide bond presenting qualitative homogeneity, while quantitative analysis suggested that the lock-release method offers more control over the conjugation process. DLS revealed differences in stability between the native and the conjugated samples at increased temperatures, with the first presenting minor changes in size and greater inherent stability. ADCs displayed irreversible structure expansion, due to partial reduction of their inter-chain disulfide bonds prior to conjugation, while the narrower distribution of the ADC (LR) at the end of the experiment is probably indicative of the lock-release method resulting in higher purity compared to the solution method. mAb (S) appeared to be the purest of all four due to its smallest size both at the beginning and at the end of the experiment. Aggregation was not evident up to 75 °C, the highest temperature at which DLS was conducted. Similarly, CD indicated differences in structure at 20 °C with mAb (S) presenting slightly stronger intensity and ADC (S) weaker signal compared to the rest. However all samples appeared unchanged at least up to 54 °C, while stability differences were emphasised at increased temperatures, as the ADCs started changing much earlier than the mAbs (Figure 3.31) and unfolded at 80 °C, while the mAbs unfolded at 82 °C. During the temperature experiment, a decrease in the β-sheet content was noted and an increase in random coil (other). The spectra after cooling down to 20 °C indicated a mixture of irreversible unfolding and aggregation.

In conclusion, temperature dependence measurements are of major help in comparability studies and especially in the case of antibodies, where despite the absence of inter-chain disulfide bonds, noncovalent interactions between opposing domains are still sufficient to maintain the compact structure under normal conditions. However, at increased temperatures, ADCs presented larger flexibility of the hinge region and unfolded earlier than mAbs, while the presence of the toxin might have also contributed to decreased stability.

References

1. K. Strebhardt and A. Ullrich, *Nature Reviews Cancer*, 2008, **8**, 473-480.
2. G. Kohler and C. Milstein, *Nature*, 1975, **256**, 495-497.
3. Ortho multicenter transplant study group, The *New England Journal of Medicine*, 1985, **313**, 337-342.

4. S. L. Morrison, M. J. Johnson, L. A. Herzenberg and V. T. Oi, *Proceedings of the National Academy of Sciences*, 1984, **81**, 6851-6855.
5. P. T. Jones, P. H. Dear, J. Foote, M. S. Neuberger and G. Winter, *Nature*, 1986, **321**, 522-525.
6. J. McCafferty, A. D. Griffiths, G. Winter and D. J. Chiswell, *Nature*, 1990, **348**, 552.
7. T. Jostock, M. Vanhove, E. Brepoels, R. van Gool, M. Daukandt, A. Wehnert, R. van Hegelsom, D. Dransfield, D. Sexton, M. Devlin, A. Ley and J. Müllberg, *Journal of Immunological Methods*, 2004, **289**, 65-80.
8. G. Winter, A. D. Griffiths, R. E. Hawkins and H. R. Hoogenboom, *Annual Review of Immunology*, 1994, **12**, 433-455.
9. N. Lonberg, L. D. Taylor, F. A. Harding, M. Trounstine, K. M. Higgins, S. R. Schramm, C. C. Kuo, R. Mashayekh, K. Wymore and J. G. McCabe, *Nature*, 1994, **368**, 856.
10. L. L. Green, M. C. Hardy, C. E. Maynard-Currie, H. Tsuda, D. M. Louie, M. J. Mendez, H. Abderrahim, M. Noguchi, D. H. Smith, Y. Zeng, N. E. David, H. Sasai, D. Garza, D. G. Brenner, J. F. Hales, R. P. McGuinness, D. J. Capon, S. Klapholz and A. Jakobovits, *Nature Genetics*, 1994, **7**, 13-21.
11. M. K. Wolpert-Defilippes, R. H. Adamson, R. L. Cysyk and D. G. Johns, *Biochemical Pharmacology*, 1975, **24**, 751-754.
12. S. Remillard, L. Rebhun, G. Howie and S. Kupchan, *Science*, 1975, **189**, 1002-1005.
13. R. Bai, G. R. Petit and E. Hamel, *Biochemical Pharmacology*, 1990, **39**, 1941-1949.
14. N. Zein, A. M. Sinha, W. J. McGahren and G. A. Ellestad, *Science*, 1988, **240**, 1198.
15. L. Ducry and B. Stump, *Bioconjugate Chemistry*, 2010, **21**, 5-13.
16. European Medicines Agency, Refusal assessment report for Mylotarg, http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/000705/WC500070677.pdf, (accessed 1 Apr, 2017).
17. U.S. Food and Drug Administration, Mylotarg (gemtuzumab ozogamicin): Market Withdrawal, https://wayback.archive-it.org/7993/20170112165432/http://www.fda.gov/Safety/MedWatch/SafetyInformation/SafetyAlertsforHumanMedicalProducts/ucm216458.htm, (accessed 1 Apr, 2017).
18. European Medicines Agency, CHMP summary of positive opinion for Adcetris, http://www.ema.europa.eu/docs/en_GB/document_library/Summary_of_opinion_-_Initial_authorisation/human/002455/WC500130143.pdf, (accessed 1 Apr, 2017).
19. P. D. Senter and E. L. Sievers, *Nature Biotechnology*, 2012, **30**, 631-637.
20. European Medicines Agency, CHMP summary of positive opinion for Kadcyla, http://www.ema.europa.eu/docs/en_GB/document_library/Summary_of_opinion_-_Initial_authorisation/human/002389/WC500150135.pdf, (accessed 1 Apr, 2017).

21. European Medicines Agency, Date of issue of marketing authorisation valid throughout the European Union for Herceptin, http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/medicines/000278/human_med_000818.jsp&mid=WC0b01ac058001d124, (accessed 1 Apr, 2017).
22. L. W. Dick, D. Qiu, D. Mahon, M. Adamo and K. C. Cheng, *Biotechnology and Bioengineering*, 2008, **100**, 1132-1143.
23. E. T. J. van den Bremer, F. J. Beurskens, M. Voorhorst, P. J. Engelberts, R. N. de Jong, B. G. van der Boom, E. M. Cook, M. A. Lindorfer, R. P. Taylor, P. H. C. van Berkel and P. W. H. I. Parren, *mAbs*, 2015, **7**, 672-680.
24. A. Ebens, F. Jacobson, P. Polakis, R. Schwall, M. Sliwkowski and S. Spencer. Antibody-drug conjugates and methods. 2005.
25. D. J. Evans and C. M. Mckee. Method of synthesising adcs using affinity resins. 2014.
26. European Medicines Agency, ICH Topic Q 6 B Specifications: Test Procedures and Acceptance Criteria for Biotechnological/Biological Products http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002824.pdf, (accessed 4 Apr, 2017).
27. European Medicines Agency, Herceptin: EPAR - Product Information, http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Product_Information/human/000278/WC500074922.pdf, (accessed 2 Apr, 2017).
28. P. M. LoRusso, D. Weiss, E. Guardino, S. Girish and M. X. Sliwkowski, *Clinical Cancer Research*, 2011, **17**, 6437-6447.
29. European Medicines Agency, Kadcyla: EPAR - Product Information, http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Product_Information/human/002389/WC500158593.pdf, (accessed 2 Apr, 2017).
30. L. J. Harris, S. B. Larson, K. W. Hasel and A. McPherson, *Biochemistry*, 1997, **36**, 1581-1597.
31. N. Guex and M. C. Peitsch, *Electrophoresis*, 1997, **18**, 2714-2723.
32. E. E. Hong and R. Chari, in *Antibody-Drug Conjugates: The 21st Century Magic Bullets for Cancer*, eds. J. Wang, W.-C. Shen and J. L. Zaro, Springer International Publishing, Cham, 2015, DOI: 10.1007/978-3-319-13081-1_4, pp. 49-76.
33. K. J. Hamblett, P. D. Senter, D. F. Chace, M. M. C. Sun, J. Lenox, C. G. Cerveny, K. M. Kissler, S. X. Bernhardt, A. K. Kopcha, R. F. Zabinski, D. L. Meyer and J. A. Francisco, *Clinical Cancer Research*, 2004, **10**, 7063-7070.
34. C. F. McDonagh, E. Turcott, L. Westendorf, J. B. Webster, S. C. Alley, K. Kim, J. Andreyka, I. Stone, K. J. Hamblett, J. A. Francisco and P. Carter, *Protein Engineering, Design and Selection*, 2006, **19**, 299-307.
35. N. Takahashi, I. Ishii, H. Ishihara, M. Mori, S. Tejima, R. Jefferis, S. Endo and Y. Arata, *Biochemistry*, 1987, **26**, 1137-1144.
36. H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore and J. Darnell, *Protein Glycosylation in the ER and Golgi Complex. In: Molecular Cell Biology. 4th edition. New York: W. H. Freeman.*, 2000.

37. J. E. Nettleship, in *Glycosylation*, ed. S. Petrescu, InTech, Rijeka, 2012, DOI: 10.5772/48154, p. Ch. 03.
38. R. Jefferis, *Biotechnology Progress*, 2005, **21**, 11-16.
39. R. Jefferis and J. Lund, *Immunology Letters*, 2002, **82**, 57-65.
40. D. F. Wyss and G. Wagner, *Current Opinion in Biotechnology*, 1996, **7**, 409-416.
41. J. N. Arnold, M. R. Wormald, R. B. Sim, P. M. Rudd and R. A. Dwek, *Annual Review of Immunology*, 2007, **25**, 21-50.
42. D. Reusch and M. L. Tejada, *Glycobiology*, 2015, **25**, 1325-1334.
43. R. J. Harris, B. Kabakoff, F. D. Macchi, F. J. Shen, M. Kwong, J. D. Andya, S. J. Shire, N. Bjork, K. Totpal and A. B. Chen, *Journal of Chromatography B: Biomedical Sciences and Applications*, 2001, **752**, 233-245.
44. K. Diepold, K. Bomans, M. Wiedmann, B. Zimmermann, A. Petzold, T. Schlothauer, R. Mueller, B. Moritz, J. O. Stracke, M. Mølhøj, D. Reusch and P. Bulau, *PLOS ONE*, 2012, **7**, e30295.
45. D. Chelius, D. S. Rehder and P. V. Bondarenko, *Analytical Chemistry*, 2005, **77**, 6004-6011.
46. A. R. Friedman, A. K. Ichhpurani, D. M. Brown, R. M. Hillman, L. F. Krabill, R. A. Martin, H. A. Zurcher-Neely and D. M. Guido, *International Journal of Peptide and Protein Research*, 1991, **37**, 14-20.
47. G. W. Becker, P. M. Tackitt, W. W. Bromer, D. S. Lefeber and R. M. Riggin, *Biotechnology and Applied Biochemistry*, 1988, **10**, 326-337.
48. N. P. Sargaeva, C. Lin and P. B. O'Connor, *Analytical Chemistry*, 2009, **81**, 9778-9786.
49. C. M. Eakin, A. Miller, J. Kerr, J. Kung and A. Wallace, *Frontiers in Pharmacology*, 2014, **5**, 87.
50. M. Hensel, R. Steurer, J. Fichtl, C. Elger, F. Wedekind, A. Petzold, T. Schlothauer, M. Molhoj, D. Reusch and P. Bulau, *PLOS ONE*, 2011, **6**, e17708.
51. Z. Wei, J. Feng, H. Y. Lin, S. Mullapudi, E. Bishop, G. I. Tous, J. Casas-Finet, F. Hakki, R. Strouse and M. A. Schenerman, *Analytical Chemistry*, 2007, **79**, 2797-2805.
52. F. J. Shen, M. Y. Kwong, R. G. Keck and R. J. Harris, *Techniques in Protein Chemistry*, 1996, **7**, 275-284.
53. M. Amano, N. Kobayashi, M. Yabuta, S. Uchiyama and K. Fukui, *Analytical Chemistry*, 2014, **86**, 7536-7543.
54. T. Zhang, J. Zhang, D. Hewitt, B. Tran, X. Gao, Z. J. Qiu, M. Tejada, H. Gazzano-Santoro and Y.-H. Kao, *Analytical Chemistry*, 2012, **84**, 7112-7123.
55. H. Pan, K. Chen, L. Chu, F. Kinderman, I. Apostol and G. Huang, *Protein Science*, 2009, **18**, 424-433.
56. B. Ghesquière, V. Jonckheere, N. Colaert, J. Van Durme, E. Timmerman, M. Goethals, J. Schymkowitz, F. Rousseau, J. Vandekerckhove and K. Gevaert, *Molecular & Cellular Proteomics : MCP*, 2011, **10**, M110.006866.
57. B. Feng, Z. Wang, T. Liu, R. Jin, S. Wang, W. Wang, G. Xiao and Z. Zhou, *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 2014, **1842**, 2345-2356.

58. E. R. Stadtman, H. Van Remmen, A. Richardson, N. B. Wehr and R. L. Levine, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 2005, **1703**, 135-140.
59. M. Haberger, A. K. Heidenreich, T. Schlothauer, M. Hook, J. Gassner, K. Bomans, M. Yegres, A. Zwick, B. Zimmermann, H. Wegele, L. Bonnington, D. Reusch and P. Bulau, *mAbs*, 2015, **7**, 891-900.
60. N. P. Chmel, S. E. Howson, L. E. N. Allan, J. Barker, G. J. Clarkson, S. S. Turner and P. Scott, *Dalton Transactions*, 2010, **39**, 2919-2927.
61. A. Damianoglou, E. J. Crust, M. R. Hicks, S. E. Howson, A. E. Knight, J. Ravi, P. Scott and A. Rodger, *Chirality*, 2008, **20**, 1029-1038.
62. V. Hall, A. Nash and A. Rodger, *Analytical Methods*, 2014, **6**, 6721-6726.
63. E. L. Gilroy, M. R. Hicks, D. J. Smith and A. Rodger, *Analyst*, 2011, **136**, 4159-4163.
64. Worldwide Malvern Instruments, Dynamic light scattering common terms defined, http://www.biophysics.bioc.cam.ac.uk/wp-content/uploads/2011/02/DLS_Terms_defined_Malvern.pdf, (accessed 2 Apr, 2017).
65. E. Castiglioni, P. Albertini and S. Abbate, *Chirality*, 2010, **22**, E142-E148.
66. F. Z. Chen, D. L. Judge, C. Y. R. Wu and J. Caldwell, *Planetary and Space Science*, 1998, **47**, 261-266.
67. T. Inagaki, *The Journal of Chemical Physics*, 1972, **57**, 2526-2530.
68. B. A. Wallace and R. W. Janes, *Modern Techniques for Circular Dichroism and Synchrotron Radiation Circular Dichroism Spectroscopy*, IOS Press, 2009.
69. S. M. Kelly, T. J. Jess and N. C. Price, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 2005, **1751**, 119-139.
70. N. P. Chmel, P. Scott and A. Rodger, *Chirality*, 2012, **24**, 699-705.
71. A. J. Miles, F. Wien, J. G. Lees, A. Rodger, R. W. Janes and B. A. Wallace, *Spectroscopy*, 2003, **17**.
72. A. J. Miles, F. Wien, J. G. Lees and B. Wallace, *Spectroscopy*, 2005, **19**, 43-51.
73. L. Chen, L. Wang, H. Shion, C. Yu, Y. Q. Yu, L. Zhu, M. Li, W. Chen and K. Gao, *mAbs*, 2016, **8**, 1210-1223.
74. R. Gahoual, A. Burr, J.-M. Busnel, L. Kuhn, P. Hammann, A. Beck, Y.-N. François and E. Leize-Wagner, *mAbs*, 2013, **5**, 479-490.
75. A. Beck, S. Sanglier-Cianférani and A. Van Dorsselaer, *Analytical Chemistry*, 2012, **84**, 4637-4646.
76. C. W. N. Damen, W. Chen, A. B. Chakraborty, M. van Oosterhout, J. R. Mazzeo, J. C. Gebler, J. H. M. Schellens, H. Rosing and J. H. Beijnen, *Journal of the American Society for Mass Spectrometry*, 2009, **20**, 2021-2033.
77. H. Yang and R. A. Zubarev, *Electrophoresis*, 2010, **31**, 1764-1772.
78. Y. W. Lao, M. Gungormusler-Yilmaz, S. Shuvo, T. Verbeke, V. Spicer and O. V. Krokhin, *Journal of Proteomics*, 2015, **125**, 131-139.
79. R. M. Pabari, B. Ryan, W. Ahmad and Z. Ramtoola, *Current Pharmaceutical Biotechnology*, 2013, **14**, 220-225.
80. M. Paul, V. Vieillard, R. Da Silva Lemos, L. Escalup and A. Astier, *International Journal of Pharmaceutics*, 2013, **448**, 101-104.
81. A. W. P. Vermeer and W. Norde, *Biophysical Journal*, 2000, **78**, 394-404.

82. R. M. Ionescu, J. Vlasak, C. Price and M. Kirchmeier, *Journal of Pharmaceutical Sciences*, 2008, **97**, 1414-1426.

Abstract

For the past 20 years, the imminent patent expiration for many marketed biopharmaceutical products has led to substantial investments, of pharmaceutical companies, in the development of biosimilar products. However, such products need to be extensively tested for their similarity to the original drug before they are granted approval by medicine organisations. Physicochemical characterisation is one of the required tests, using a variety of analytical methods in order to provide sufficient data on the biosimilar compared to the reference product as well as on batch-to-batch comparison. In the present work, real cases of five therapeutic proteins were examined and are presented as short studies: protein H, somatropin, protein A, etanercept and a monoclonal antibody. The main purpose of all studies was comparability assessment of the secondary structure of the provided samples using far-UV CD spectroscopy. In most cases, the CD spectra were analysed for their secondary structure content using SELCON3 and CDSSTR with various reference sets, and SSNN. A variety of analytical techniques were used in different combinations in each study in order to further complement far-UV CD data. FT-IR provided supporting information on secondary structure, DLS served for size comparability, Near-UV CD and fluorescence depends on tertiary structure and orientation of the aromatic amino acids of the under-analysis protein, and 1D ^{1}H-NMR provided higher-order structure comparison. Coupling of CD and/or DLS with temperature-dependent measurements provided information on the thermal stability of the samples. The encountered challenges, resulting mostly from high concentration of the samples and the formulation buffer, were addressed in the best possible way given the limited amount of time in which the studies had to be completed.

Chapter 4
Real Cases of Biotherapeutic Proteins: Short Studies and Challenges

4.1. INTRODUCTION

The examined cases reported in this chapter concern protein therapeutics that have long been available in the market, namely protein H,* somatropin (hGH), protein A, etanercept (TNFR2:Fc) and a monoclonal antibody (mAb). They are short studies on real cases of biosimilar products which are currently under development. Our aim was to provide a comparability assessment between different batches (for protein A, etanercept, mAb), between the biosimilar and the reference product (for etanercept) or between the native and the conjugated antibody (for protein H). Samples were provided by BioPharmaSpec and assessed for their particle size, secondary/tertiary structure, aromatic amino acids environment and higher-order structure similarity according to the manufacturer's requests, or out of additional research interest. For these purposes, the applied techniques were DLS, far-UV CD and FT-IR, near-UV CD, fluorescence and 1D ^{1}H-NMR respectively, used in different combinations for each study. In some cases, thermal stability was also tested by coupling CD and/or DLS with temperature-dependent measurements (etanercept and mAb).

Far-UV CD was the common technique in all studies, as secondary structure comparability was the main request in all examined cases. For the secondary structure analysis of the far-UV CD spectra, several of the available algorithms were used and their results were compared with the DSSP assignment for the crystal structure (from suitable PDB entries). However, several challenges were encountered, mostly occurring due to high viscosity of the concentrated samples. Dilution of the provided samples was necessary in certain cases (*e.g.* mAb study), where the protein concentration was extremely high and the formulation buffer was not suitable for a particular technique (such as CD and NMR). Dilution is not optimal as this might induce changes in the secondary structure of the examined molecule (as was the case

* Data on protein H and protein A are confidential and therefore excluded from the present Volume. A separate Volume (Vol. 2) is provided for the examiners to review.

with protein A). The best possible solutions were provided, given the limited amount of time in which these studies had to be completed. In most of the cases measurements were conducted in three replicates in order to ensure reproducibility of the results.

4.1.1. Protein H

Protein H concerns confidential material, therefore any information regarding the protein name and structure has been excluded from the main thesis (Vol. 1). The protein is introduced and discussed in a separate Volume (Vol. 2), which is provided for the examiners to review.

4.1.2. Somatropin (human growth hormone, hGH)

Human growth hormone is responsible for growth-promoting effects in the body, and its secretion from the anterior pituitary gland is controlled by the hypothalamus.[1] The synthetic form of the protein is called somatropin (recombinant human growth hormone rhGH; here referred to as hGH for simplicity).[2,3] Human growth hormone is essential for normal growth in children, by acting directly on the growth plates or by production of insulin-like growth factors (*e.g.* IGF-1) mainly in the liver.[4] It also plays an important role in the metabolism of proteins, lipids and carbohydrates during childhood and throughout adult life. Humatrope (by Lilly) was the first somatropin product to receive FDA approval in 1987 for treatment of children with growth failure associated with a number of causes such as inadequate endogenous growth hormone secretion (growth hormone deficiency, GHD) and Turner syndrome (TS), and for replacement of endogenous growth hormone in adults with GHD. Numerous commercial products of somatropin have become available since, with slight differences in the indications, while Omnitrope (by Sandoz GmbH) was the first biosimilar to gain marketing-authorisation approval by EMA in 2006.

However, the fact that most of the currently established products require a once daily subcutaneous injection makes adherence to the treatment troublesome, especially for paediatric patients. Efforts are being made to develop long-acting, sustained-release and prolonged half-life preparations that will allow for reduced injection frequency and decreased inconvenience of a daily therapy. Such technologies include microencapsulation of rhGH in microparticles and protein fusion/conjugation as follows:

1. Microparticles consisting of growth hormone incorporated into a hyaluronidate matrix (LG Life Sciences and BioPartners)[5] result in sustained-release of rhGH.

2. Fusion of a sequence of hydrophilic amino acids as head and tail to both C- and N-termini of rhGH results in a MW of 119 kDa, >5 times rhGH's mass (XTEN by Amunix).[6] Increase of the hydrodynamic size of the molecule results in reduced clearance by filtration in the kidneys (glomerular filtration), therefore extending the half-life of rhGH.
3. Conjugation of a carboxyl-terminus peptide (CTP) from human chorionic gonadotropin (hCG) to rhGH (Prolor Biotech Inc.)[7] prolongs half-life.
4. The inactive prodrug consisting of PEGylated rhGH (PEG = polyethyleneglycol) linked to a carrier (TransCon by Ascendis Pharma)[8] releases GH in its native form through hydrolysis of the linker depending on pH and temperature conditions, resulting in a sustained-release prodrug.
5. Fusion of human serum albumin with the N-terminus of rhGH (Albutropin by TEVA)[9] or via a terminal fatty acid, which is attached to the rhGH backbone and exhibits noncovalent albumin-binding properties (reversible albumin binding by Novo Nordisk),[10] both extend half-life of the product.
6. A hybrid consisting of rhGH fused with a hybrid Fc fragment containing partial Fc domains of human IgD and IgG4 produces a long-acting fusion protein (LAPSCOVERY by Hanmi).[11]

These novel formulations aim at once per week or less frequent injections, maintaining effective drug levels, equivalent to those obtained with daily hGH therapy, throughout the dosing interval. Moreover, the levels of both hGH and IGF-1 should neither exceed nor fall below the "normal" ones, as both excess (acromegaly) and deficiency (GHD) are related with increased rates of morbidity and mortality due to consequences in the cardiovascular system.[12,13] Tolerability of the induced modifications and maintaining the original activity and biodistribution of hGH are essential for achieving optimal efficacy and safety, while at the same time the manufacturing and distribution costs need to be kept within affordable levels. Nutropin Depot (Genentech) was the first long-acting formulation to receive FDA-approval in 1999, using zinc-stabilised rhGH embedded in PLGA (polylactic-co-glycolic acid) microspheres (ProLease by Alkermes Inc.),[14] however, it was withdrawn from the market in 2004 due to the high cost of its production.[15]

Structure

Human growth hormone consists of a single polypeptide chain of 191 amino acids resulting in a molecular mass of ~22 kDa (Figure 4.1). Two intra-molecular disulfide

bonds are formed between Cys53–Cys165 and Cys182–Cys189 (Figure 4.2 in yellow). Previous studies have shown that hGH is a globular protein in solution,[16,17] compactly folded and presenting a hydrophobic interior. X-Ray crystallography has revealed a mainly α-helical conformation of approximately 52%[18] which is distributed into four helical segments, while the helix content from CD spectra has been calculated as 55±5%.[19] The rigid helices are believed to be somewhat conserved among different species, while the surrounding loops are probably more flexible and responsible for species-specific properties.[20]

Somatropin – Average Mass = 22129.05 Da

```
1   FPTIP LSRLF DNAML RAHRL HQLAF DTYQE FEEAY IPKEQ KYSFL QNPQT
51  SLCFS ESIPT PSNRE ETQQK SNLEL LRISL LLIQS WLEPV QFLRS VFANS
101 LVYGA SDSNV YDLLK DLEEG IQTLM GRLED GSPRT GQIFK QTYSK FDTNS
151 HNDDA LLKNY GLLYC FRKDM DKVET FLRIV QCRSV EGSCG F
```

Figure 4.1: Primary structure of human growth hormone, consisting of a single polypeptide chain of 191 amino acids and a total mass of ~22 kDa. The cysteine residues which contribute to the formation of the intra-chain disulfide bonds are highlighted in yellow.

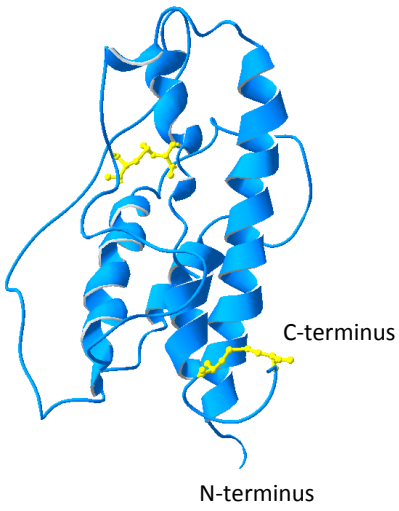


Figure 4.2: Three-dimensional structure of hGH (PDB ID 1HGU)[21] in ribbon representation. The shown model has 12 mutations and therefore might vary slightly from the native hGH. The side chains of cysteine residues, which contribute to the formation of the intra-molecular disulfide bridges, are shown in ball-and-stick mode (in yellow).

4.1.3. Protein A

Protein A concerns confidential material, therefore any information regarding the protein name and structure has been excluded from the main thesis (Vol. 1). The protein is introduced and discussed in a separate Volume (Vol. 2), which is provided for the examiners to review.

4.1.4. Etanercept

Etanercept is a therapeutic protein for regulation of various inflammatory and autoimmune diseases, and finds medical use in the treatment of rheumatoid arthritis (RA).[22] It is a dimeric fusion protein produced by recombinant DNA engineering and consists of the extracellular portion of the human 75 kDa (p75) tumour necrosis factor receptor (TNFR2) linked to the Fc component of human immunoglobulin G1 (IgG1). Etanercept belongs to the wider class of tumour necrosis factor antagonists that are commercially available, along with the anti-TNF mAbs adalimumab and infliximab. Its mechanism of action comprises neutralisation of the tumour necrosis factor (a soluble inflammatory cytokine) by inhibiting binding of TNFα to cell surface TNF receptors and therefore by disrupting inflammatory pathways driven by TNF.[23] Etanercept was the first anti-TNF biological agent to receive approval by FDA in 1998 and by EMA in 2000 under the trade name Enbrel, and was indicated for reduction in symptoms of active rheumatoid arthritis in patients who have had an inadequate response to disease-modifying antirheumatic drugs (DMARDs).[24,25]

Antibody-based therapeutic proteins have the advantage of binding to the target with high specificity, and therefore being more effective and presenting fewer side effects compared to conventional DMARDs. Moreover, their extended half-life in the blood circulation allows for less frequent administration. In the case of etanercept, the half-life of TNFR (30 min–2 h)[26] is extended to 70–132 h[27] due to the Fc portion of IgG.[28] Enbrel is administered by subcutaneous injection and is provided as prefilled syringes containing 50 mg/mL solution of etanercept and formulated with: sucrose (10 mg/mL), sodium chloride NaCl (5.8 mg/mL), L-arginine HCl (5.3 mg/mL), sodium phosphate monobasic monohydrate (2.6 mg/mL) and sodium phosphate dibasic anhydrous (0.9 mg/mL) at pH 6.3. Vials containing 25 mg of etanercept powder for reconstitution are formulated with: sucrose (10 mg), mannitol (40 mg) and tromethamine (1.2 mg) at pH 7.4.

Biopharmaceutical companies have been making efforts to develop Enbrel biosimilars in order to reduce the cost of RA therapy.[29] However, the complexity of the molecule and especially glycosylation patterns render biosimilar development challenging.[30] Among other biosimilars that have been approved for inflammatory diseases,[31] two biosimilars of Enbrel were approved in 2016: Benepali by EMA[32] and Erelzi by FDA,[33] and one more in 2017: Lifmior by EMA.[34] The present study investigates the similarity between 3 biosimilar and 3 innovator products of etanercept, from 3 different batches in each case. DLS was used in order to provide a size and

stability comparability assessment of the provided samples, far-UV CD and FT-IR provided information on secondary structure similarity and stability, near-UV CD assessed the tertiary structure of the products and fluorescence evaluated similarity in the environment of the aromatic amino acids.

Structure

As described above, etanercept is a homodimer and each monomer (or chain) contains the TNFR2 molecule linked to the Fc portion of IgG1 by an O-glycosylated peptide.[35] Each chain consists of 467 amino acids and has a molecular mass of ~51 kDa (Figure 4.3), while the two chains are held covalently via 3 inter-chain disulfide bonds at Cys (240), Cys (246) and Cys (249), resulting in a total molecular mass of ~102 kDa. The rest of the cysteine residues contribute to the formation of 13 intra-chain disulfide bonds (on each chain), 11 of which are in the TNFR2 region, and 2 in the Fc region.[36] Etanercept presents extensive glycosylation, which is the main cause for its heterogeneity, especially for the O-linked glycans. Each chain has 3 N-glycosylation sites: Asn (149), Asn (171) on the TNFR2 and Asn (317) in the Fc region, and 10–13 potential O-glycosylation sites, resulting in a mass of up to 130 kDa for the glycosylated molecule (65 kDa for each chain).[30,37] Figure 4.3 shows the TNFR2 region, the O-glycosylated linker and the Fc portion underlined in green, red and blue respectively. The cysteine residues that form inter-chain disulfide bridges are highlighted in yellow and the glycosylation sites are highlighted in orange. It has been reported that one more O-glycosylation site exists in the region 121–185 of the sequence.[35]

Etanercept – Average Mass = 51234.90 Da

```
1    LPAQV AFTPY APEPG STCRL REYYD QTAQM CCSKC SPGQH AKVFC TKTSD
51   TVCDS CEDST YTQLW NWVPE CLSCG SRCSS DQVET QACTR EQNRI CTCRP
101  GWYCA LSKQE GCRLC APLRK CRPGF GVARP GTETS DVVCK PCAPG TFSNT
151  TSSTD ICRPH QICNV VAIPG NASMD AVCTS TSPTR SMAPG AVHLP QPVST
201  RSQHT QPTPE PSTAP STSFL LPMGP SPPAE GSTGD EPKSC DKTHT CPPCP
251  APELL GGPSV FLFPP KPKDT LMISR TPEVT CVVVD VSHED PEVKF NWYVD
301  GVEVH NAKTK PREEQ YNSTY RVVSV LTVLH QDWLN GKEYK CKVSN KALPA
351  PIEKT ISKAK GQPRE PQVYT LPPSR EEMTK NQVSL TCLVK GFYPS DIAVE
401  WESNG QPENN YKTTP PVLDS DGSFF LYSKL TVDKS RWQQG NVFSC SVMHE
451  ALHNH YTQKS LSLSP GK
```

Figure 4.3: Primary structure of a single chain of etanercept. The sequence of the TNFR2 molecule is underlined in green, the O-glycosylated linker with red and the Fc portion of IgG1 with blue. The cysteine residues that contribute to the formation of inter-chain disulfide bridges are highlighted in yellow and the potential N- and O-glycosylation sites in orange.

Since there is no crystal structure available for the whole molecule in the PDB, the secondary structure of etanercept was calculated by combination of the PDB entries for TNFR2 (PDB ID: 3ALQ)[38] and the Fc fragment of IgG1 (PDB ID: 3AVE),[39] and

showed ~33% β-sheet, ~17% turns, ~5% α-helix, while structure was not assigned for ~45% of the amino acids in the protein. The structure in Figure 4.4 shows the 4 cysteine-rich domains of TNFR2 (orange) and the Fc fragment of IgG1 consisting of the hinge region with the 3 inter-chain disulfide bonds, C_H2 and C_H3 domains (blue) and the intermediate O-glycosylated peptide.

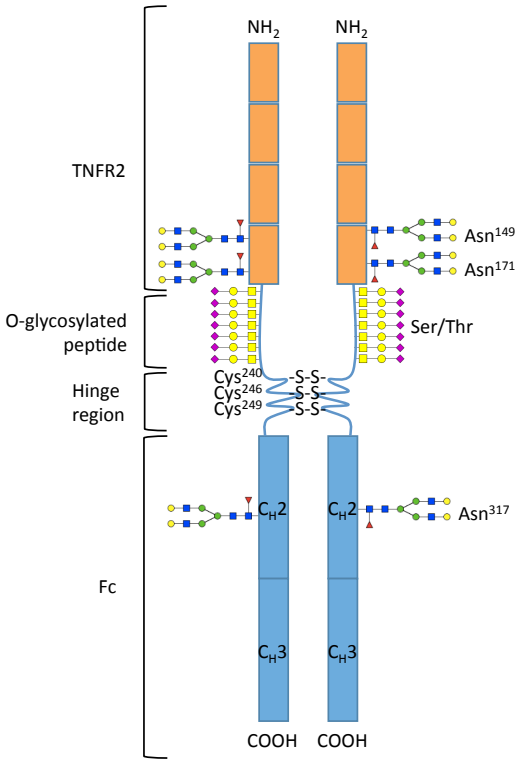


Figure 4.4: Schematic representation of etanercept structure showing the four domains of TNFR2 (orange) and the C_H2 and C_H3 domains of the Fc fragment of IgG1 (blue). Glycosylation in etanercept takes place on the asparagine residues Asn (149), Asn (171) and Asn (317), while 10–13 potential O-glycosylation sites have been reported (here only a few of them are shown in the O-glycosylated linker region). Only the inter-chain disulfide bridges of the hinge region are shown (-S-S-). The glycan structures were built with GlycanBuilder[40,41]

4.1.5. Monoclonal antibody (mAb)

Development of suitable formulations for protein therapeutics aims at providing long-term storage stability by protecting the protein from degradation (such as oxidation and deamidation) and minimising protein-protein interactions, which could potentially result in formation of immunogenic aggregates. More specifically, therapeutic monoclonal antibodies are often required to be formulated in high concentrations (~100 mg/mL or even more)[42] in order to facilitate administration and reduce the injected volume of the usually high dosage administered. However, aggregation and high viscosity are two of the major concerns in concentrated antibody formulations. In order to tackle these complications and to minimise aggregation, addition of excipients

such as surfactants and stabilisers, and adjustment of the pH conditions in a range far from the isoelectric point (pI) of the protein are fundamental actions during formulation development.[43,44] Moreover, such high protein concentrations induce analytical challenges, as most of the quality control tests normally require lower concentrations and therefore dilution of the product is necessary, though not ideal as the dilution process may not be reproducible and can affect the protein structure. Different concentrations can presumably result in structural and stability changes, however given that all samples are treated in a similar way, this work is predicated on the assumption that a comparability assessment is still allowed within each one of the used analytical methods.

In this study, four monoclonal antibody samples were provided in their final formulation, at a concentration of ~100 mg/mL, with the objective of investigating their similarity. The excipients used in this case were Tween 80 as surfactant, arginine as stabiliser and the pH was adjusted to 5.0 as usually recommended for antibodies, which are stable at pH 4.5–5.5. Dilution or buffer exchange of the samples was necessary prior to the chracacterisation experiments, due to their high concentration. DLS was used in order to provide a size comparability assessment of the samples, far-UV CD provided information on secondary structure similarity and thermal stability and 1D ^{1}H-NMR assessed similarity in the higher-order structure of the products.

4.2. MATERIALS AND METHODS

4.2.1. Materials

All materials used for the following studies were as described in Chapter 2. For the CD calibration and path length assignment the same apply as described in materials of Chapter 3. All samples and buffers were provided by BioPharmaSpec Ltd. as follows:

Table 4.1: Somatropin (hGH) sample and formulation buffer as received from BioPharmaSpec Ltd. The buffer was 20 mM ammonium bicarbonate, pH 8.3.

Sample name in text	Sample ID	Protein Concentration (mg/mL)	Volume supplied (mL)
Buffer	Buffer	–	5
hGH	Somatropin	3.56	0.250

Table 4.2: Etanercept (TNFR2:Fc) samples and buffer as received from BioPharmaSpec Ltd. The colours are according to the figures in the results section (Section 4.6). For the 1st analysis the samples were provided as buffer exchanged into 50 mM ammonium bicarbonate buffer pH 7.07. For the 2nd analysis (requested repeat analysis for the fluorescence data, Section 4.6.4) the samples were in a formulation consisting of: sucrose 10 mg/mL, NaCl 5.84 mg/mL, L-Arginine HCl 5.3 mg/mL, sodium phosphate monobasic monohydrate 2.4 mg/mL and sodium phosphate dibasic anhydrous 2.03 mg/mL; PBS buffer (Sigma P4417) was supplied as a diluent.

		1st Analysis		2nd Analysis	
Sample name in text	Sample ID	Protein Concentration (mg/mL)	Volume supplied (mL)	Protein Concentration (mg/mL)	Volume supplied (mL)
TNFR2:Fc 1	Etanercept	0.88	2	50	0.050
TNFR2:Fc 2	Etanercept	0.97	2	50	0.050
TNFR2:Fc 3	Etanercept	0.89	2	50	0.050
TNFR2:Fc 4	Etanercept	0.88	2	25	0.100
TNFR2:Fc 5	Etanercept	0.97	2	25	0.100
TNFR2:Fc 6	Etanercept	0.95	2	25	0.100
Buffer	Buffer	–	50	–	100

Table 4.3: Monoclonal antibody (mAb) samples and buffers as received from BioPharmaSpec Ltd. The colours are according to the figures in the results section (Section 4.7). The samples were in a formulation consisting of: 50 mM citric acid monohydrate, 0.05% w/v Tween 80, 110 mM L-Arginine, 240 mM L-Arginine HCl, at pH 5.0; a similar buffer but without Tween 80 was supplied as a diluent.

Sample name in text	Sample ID	Protein Concentration (mg/mL)	Volume supplied (mL)
mAb 1	Monoclonal Antibody	102.2	0.200
mAb 2	Monoclonal Antibody	103.0	0.200
mAb 3	Monoclonal Antibody	103.1	0.200
mAb 4	Monoclonal Antibody	96.7	0.200
Buffer	Buffer with PS80	–	50
Buffer	Buffer without PS80	–	50

4.2.2. Spectroscopic data collection

The ***CD, DLS*** and ***FT-IR data*** were collected on the instruments as described in Chapter 2 (Section 2.2.2).

For the ***CD data***, the spectrum of the reference buffer or the averaged spectrum (when the buffer was recorded in triplicate) was subtracted from each replicate. The data were zeroed by subtracting the averaged value from 260–250 nm (mdeg) or 250–240 nm (for the *monoclonal antibody*) for the ***far-UV data,*** and 350–340 nm (for *protein A*) or 330–320 nm (for *etanercept*) for the ***near-UV data*** from the whole spectrum. The ***far-UV spectra*** were converted to Δε units and averaged over 3 replicates (as presented in the text). The wavelength step was changed to 1 nm and the truncated spectra were tested using SELCON3, CDSSTR and SSNN for the secondary structure analysis, as described in Section 4.2.7.

Fluorescence spectra were recorded using a Jasco FP-6500 spectrofluorometer. For both analyses, 1st and 2nd (repeat analysis for the fluorescence data), data for *etanercept* were collected with an excitation wavelength of 280 nm and with an excitation wavelength of 295 nm. At an excitation wavelength of 280 nm, aromatic amino acids tryptophan and tyrosine fluoresce. At an excitation wavelength of 295 nm, the fluorescence intensity is dominated by tryptophan fluorescence. Samples and buffer were measured in triplicate. The mean fluorescence intensity was calculated at each wavelength for the respective buffer and subtracted from each replicate. The three subtracted replicates were averaged for each sample. There may be an element of inner filter effect reducing signals at the concentrations used in these experiments.[45,46]

For the ***DLS data*** of *etanercept* (samples from the 2nd analysis were used), the refractive index of the PBS buffer was measured as 1.334 and the viscosity of the solution was determined by using latex spheres of known diameter suspended in 10 mM sodium chloride solution and in PBS buffer, as described in Chapter 3. The size distributions are shown in Appendix B (Figure B3, Figure B4), for all measurements and averaged measurements/replicates respectively. For the particle diameters in NaCl and PBS, the Z-Average values as calculated by the Zetasizer software were used for the viscosity calculation (106.6 nm for the NaCl solution, and 107.3 nm the average value of the three replicates in PBS), resulting in a viscosity value of 1.009. Samples and buffer were measured in triplicate, and 6 measurements were averaged for each replicate. For each sample, the data are presented in terms of intensity and number values as an average over 3 replicates. The standard deviations were calculated from the three replicates of each sample, using the mean number values as calculated by the software.

For the *monoclonal antibody* samples, the refractive index of each buffer solution was measured as shown in Table 4.5 and the viscosity was determined as described in Chapter 3, by using latex spheres in 10 mM sodium chloride and in each one of the four buffer solutions. The size distributions of the latex spheres in all five solutions are shown in Appendix B (Figure B20, Figure B21) for all measurements and averaged measurements respectively. For the viscosity calculation the Z-Average values were used for the particle diameters, as shown in Appendix B (Figure B20), and the resulting viscosity values are given in Table 4.5. For each sample, the data are presented in terms of intensity and number values as an average over 8 measurements. The standard deviations were calculated from 8 measurements for each sample, using the mean number values as calculated by the software.

FT-IR spectra of *etanercept* (samples from the 1st analysis were used) were recorded in triplicate and the buffer was recorded once using a 6-bounce ZnSe Attenuated Total Reflectance flat plate with lid. The spectra were corrected by applying a factor to the buffer subtraction in order to minimise the water association peak at 2130 cm^{-1}, while a second factor was applied to the water vapour spectrum in order to minimise its contribution at 1700–1800 cm^{-1}. The spectra were zeroed in the region of 1800 cm^{-1} and averaged over the 3 replicates. The data were normalized by dividing by the maximum absorbance value between 1600 and 1800 cm^{-1} for each spectrum, and the normalised spectra were analysed for their secondary structure content in the region of the amide I band (1600–1700 cm^{-1}) using the retrained version

of SSNN (as discussed in Chapter 2). The second derivative was calculated for the area 1400–1800 cm^{-1}, using a five data point window and without smoothing.

1D ^{1}H-NMR experiments for the *monoclonal antibody* samples were carried out on a Bruker AV III-600 spectrometer operating at 600.13 MHz for ^{1}H. The water peak was suppressed using double pulsed field gradient spin echo for excitation sculpting (DPFGSE).[47] All spectra were Fourier-transformed, phased and baseline-corrected in Topspin 3.2. Exponential multiplication was used for apodisation and line broadening was 1.0 Hz. Spectra were recorded with 15624 points in the time domain and processed with 32768 points in the frequency domain. Spectral width was 13 ppm.

The ***balance, pH meter, refractometer***, ***centrifuges*** and ***water purification*** system were as described in the spectroscopic data collection of Chapter 2.

4.2.3. Measurements

For all ***far-UV CD experiments***, the same parameters were used: wavelength range 260–180 nm, 1 s data integration time (DIT), 2 nm bandwidth (unless otherwise stated), 0.2 nm data pitch, 100 nm/min continuous scanning speed. Each spectrum was the average of 16 accumulations (unless otherwise stated), at room temperature.

⇒ *Protein H samples*: Spectra of the samples (nominally 1 mg/mL) and buffer were measured in triplicate using a demountable cuvette with a nominal path length of 0.01 cm. The path length of the cuvette was measured by using the absorbance of 0.02 M potassium chromate and the Beer-Lambert Law with an extinction co-efficient of 4830 $M^{-1}cm^{-1}$ at 372 nm, and was determined to be 0.0058 ± 0.0002 cm. The value 0.0058 was used in the analysis and the mean residue weight (MRW, described in Section 1.3.2 in the CD part) was 109.6 (without accounting for the attached glycans, and for the toxin molecules in the case of the ADC).

⇒ *Somatropin (hGH) sample:* Spectra were measured in triplicate for the samples and once for the buffer. For the original concentration (nominally 3.56 mg/mL) a demountable cuvette with nominal path length of 0.001 cm was used, while spectra for the 2- and 4-fold diluted samples (df2 and df4) were obtained using a nominally 0.01 cm demountable cuvette. A cuvette with a fixed path length of 0.1 cm was used for the sample that was diluted by a factor of 30 (df30). The path lengths of the demountable cuvettes were measured by using 0.20 M and 0.02 M solutions of potassium chromate, as described above, for the 0.001 cm and 0.01 cm cuvettes respectively. The path lengths were determined as 0.000726 ± 0.000009 cm and 0.00584 ± 0.00009 cm, and the values 0.000726 and 0.00584 were used for conversion of data to Δε units. MRW was 116.5.

⇒ *Protein A samples*: For the samples in their original concentration (~2 mg/mL), spectra were measured in triplicate using a demountable cuvette with nominal path length of 0.01 cm; a 0.1 cm fixed path length cuvette was used for the 10-fold diluted samples (df10), which were recorded once. The respective buffers (in the original concentration and after 10-fold dilution) were recorded once in the appropriate cuvettes. The path length of the demountable cuvette was measured as described for protein H, and was determined to be 0.00628 ± 0.00003 cm. MRW was 123.6, using the MW of the glycosylated protein (65,000 Da) as provided with the samples. For the fixed 0.01 cm cuvette, a wavelength range of 350–180 nm was used and 8 accumulations. The path length of the fixed cuvette was accurately measured using the interference fringe method and gave a value of 0.0099 cm. This method uses interference between the incident light beam and light that has been double reflected between the two surfaces of the cuvette as described by Hennessey and Johnson.[48]

⇒ *Etanercept (TNFR2:Fc) samples*: Spectra were recorded for the 1st analysis samples at the received concentration (~0.9 mg/mL), using a demountable cuvette with a nominal path length of 0.01 cm. Both the buffer and the samples were recorded in triplicate, and each spectrum was the average of 12 accumulations. The path length of the demountable cuvette was measured as described for protein H, and was determined to be 0.0060 ± 0.0001 cm. The value 0.0060 was used for conversion to Δε units and MRW was calculated as 109.9, without taking into account glycosylation, as the purpose of the study was comparability among samples and not extracting quantitative results of the secondary structure.

⇒ *Monoclonal antibody (mAb) samples*: Spectra were recorded before (at 25 °C) and after (cooled to 25 °C) the melting curves for the 1000-fold diluted samples (~0.1 mg/mL). The buffer with Tween 80 was diluted accordingly (see sample preparation) and a 0.1 cm fixed path length cuvette was used. The same parameters were used, as described above and a wavelength range 250–180 nm.

For the ***near-UV CD experiments***, a wavelength range of 350–250 nm was used and the DIT was set to 2 s for *protein A*, while the rest of the parameters were the same as for the far-UV data. Protein A spectra were recorded once, at 37 °C. For *etanercept* (1st analysis samples) a wavelength range of 330–250 nm was used, 1.0 nm bandwidth, 16 accumulations and the rest of the parameters were as described for far-UV. Etanercept spectra were recorded in triplicate. A 1 cm path length masked black quartz cuvette was used in all cases.

Fluorescence data for *etanercept* (1st analysis samples) were recorded with emission wavelength range 280–500 nm at excitation wavelength 280 nm, and 295–500 nm at excitation wavelength 295 nm. The DIT was set to 1 s, data pitch was 0.2 nm, scanning speed was 200 nm/min and excitation and emission bandwidths were set to 3 nm. A fixed path length fluorescence, nominal 3 mm, cuvette was used. For the 2nd analysis samples, spectra were recorded with the same parameters, but 0.1 nm data pitch.

DLS measurements for *etanercept* (2nd analysis samples) were obtained as 6 measurements for each replicate, with 10 runs per measurement and a run duration of 20 s. An equilibration time of 60 s was allowed before the measurements started. For the sample measurements, material was set to protein (RI 1.450), and viscosity and refractive index of the dispersant were 1.009 and 1.334 respectively. For the latex measurements, material was set to polystyrene latex (RI 1.590) and the viscosity and refractive index of water were used. The measurements for the *monoclonal antibody* samples were obtained as 8 measurements, with 12 runs per measurement and a run duration time of 10 s. The viscosity and refractive index of the dispersant for each sample were set as shown in Table 4.5. For the buffers, the same parameters were used, but with 5 measurements and the refractive index and viscosity of water. As described for etanercept, the material was set accordingly for sample and buffer measurements. For all data a 3 mm quartz cuvette (small volume) was used.

For the ***FT-IR experiments*** for *etanercept* (1st analysis samples), the instrument was purged with N_2 for at least 1 hour before recording the first spectrum and continuously throughout the measurements. All data were collected in a wavelength range from 400–4000 cm^{-1} and for each spectrum 1000 scans with 4 cm^{-1} resolution were summed.

For the ***1D ^{1}H-NMR spectra*** the acquisition time was 1.0 s. A relaxation delay of 1.7 s, and a 90˚ pulse width of 16.00 µs were used. All data were collected at a temperature of 45 °C and the final spectra were the sum of 2560 scans.

4.2.4. Temperature dependent measurements

Far-UV CD data for the temperature-dependent measurements of *etanercept* were recorded at a concentration of 0.1 mg/ml, using a 0.1 cm fixed path length cuvette. Spectra for etanercept sample 1 (sample from the 1st analysis) were recorded in a temperature range 20–100 °C, with a 5 °C step, and a ramp rate of 2.5 °C/min, and after cooling down to 20 °C with the rest of the parameters set as described above. Melting curves were recorded for all six etanercept samples (2nd analysis) at two

different fixed wavelengths: 195.4 nm and 202.0 nm, with data points obtained from 25–90 °C, every 0.1 °C, at a ramp rate of 0.5 °C/min. The temperature was kept within ±0.1 °C from the desired temperature for 3 s before the measurement was recorded. Spectra were also collected at the beginning (25 °C) and at the end of the melting curves, after the samples were cooled down to 25 °C (parameters were same as above with the exception of a wavelength range 260–185 nm). The curves were smoothed by averaging 10 points for each temperature *e.g.* 25 °C was the average of the intensities from 25.0–25.9 °C *etc.* The average intensity of the smoothed melting curves from 25–29 °C was adjusted to the Δε values (at the respective wavelengths 195.4 nm and 200.0 nm) of the 25 °C spectra that were recorded at the beginning of each experiment.

Melting curves for the *monoclonal antibody* samples were recorded at 200.0 and 217.0 nm, with data points recorded from 25–100 °C, every 0.1 °C. The ramp rate was 0.5 °C/min and the temperature was kept within ±0.1 °C from the desired temperature for 5 s before the measurement was recorded. DIT and bandwidth were set as described in Section 4.2.3. The curves were smoothed as described for etanercept, and each curve was offset for visibility by using as a starting point the average value from 25–29 °C, averaged over the four samples.

For the ***DLS temperature experiments***, *etanercept* samples (2nd analysis) were measured from 25–75 °C with a 5 °C temperature interval. For each temperature, number of measurements, runs and run duration were as described above. An equilibration time of 180 s was allowed after the desired temperature was reached. The viscosity of the PBS buffer was estimated at each temperature as described above, and the corrected viscosity values are given in Table 4.4:

Table 4.4: Corrected values for viscosity at each temperature and refractive index for the used blank solution (PBS buffer) for the etanercept DLS study.

Temperature (°C)	25	30	35	40	45	50	55	60	65	70	75	Refractive Index
PBS buffer	0.924	0.806	0.746	0.700	0.683	0.642	0.611	0.583	0.563	0.551	0.520	1.334

4.2.5. Sample preparation

For the ***far-UV CD spectra***, the *protein H samples* were used in their received concentrations. *Somatropin* was used in its original concentration (nominally 3.56 mg/mL) and in 2-, 4- and 30-fold dilutions with water. *Protein A* was used in its original concentration (~1.9 mg/mL) and in a 10-fold dilution with water. *Etanercept* samples were used in their received concentrations (~0.9 mg/mL) for the 1st analysis. For the melting curves (2nd analysis), the samples were diluted down to 1 mg/mL with

PBS and a further 10-fold dilution (0.1 mg/mL) with water followed, as PBS is not suitable for far-UV CD measurements due to the Cl^- absorption. *Monoclonal antibody* samples were used after dilution with water and 100 mM sodium phosphate buffer (pH 7.04), resulting in a final solution of ~0.1 mg/mL protein in 10 mM sodium phosphate buffer. The buffer with Tween 80 was diluted in the same way.

For the ***near-UV CD spectra***, *protein A* samples were used in their original concentrations (~1.9 mg/mL). *Etanercept* samples and buffer were recorded as received (1st analysis, ~0.9 mg/mL).

Fluorescence data for *etanercept* were recorded in the received concentrations of the samples for the 1st analysis (~0.9 mg/mL), while for the repeat (2nd) analysis 50- or 25-fold dilution with PBS was applied, depending on the original concentration of the samples, in order to reach a final concentration of 1 mg/mL.

For the ***DLS measurements*** of *etanercept* (2nd analysis) the samples were diluted with PBS to a final concentration of 1 mg/mL. All solutions were filtered through 0.2 µM syringe filters (Thermo Scientific) directly into the quartz cuvette prior to the measurement. The *monoclonal antibody* samples were diluted to 5 mg/mL using the buffer without Tween 80. The buffer with Tween 80 was diluted similarly and the specific volumes of samples and buffers are shown in Table 4.5.

Table 4.5: Dilution of the monoclonal antibody samples to 5 mg/mL using the buffer without Tween 80. The buffer with Tween 80 was diluted according to the respective sample. Refractive indices were measured for each buffer using a refractometer, and viscosity values were calculated by suspension of latex spheres of known diameter in 10 mM NaCl solution and each of the buffers.

Buffer	Refractive Index	Viscosity (mPa)	Volume of sample/ buffer with Tween 80 (µL)	Volume of buffer without Tween 80 (µL)
Buffer 1	1.348	1.240	9.8	190.2
Buffer 2	1.348	1.236	9.7	190.3
Buffer 3	1.348	1.242	9.7	190.3
Buffer 4	1.349	1.246	10.3	189.7

FT-IR data for *etanercept* samples (1st analysis) were recorded in their received concentrations (~0.9 mg/mL).

1D ^{1}H-NMR spectra for the *monoclonal antibody* samples were recorded after buffer exchange into 20 mM sodium phosphate buffer using Amicon filters with a 30 kDa MWCO (molecular weight cut-off) and addition of 10% D_2O for the field-

frequency lock.[49] The pH of the sodium phosphate buffer was adjusted to 5.4 using 1 M hydrochloric acid solution. The data were recorded at sample concentrations of approximately 30 mg/mL (mAb 1), 48 mg/mL (mAb 2), 43 mg/mL (mAb 3) and 36 mg/mL (mAb 4).

4.2.6. CD calibration

For each study the J-1500 CD spectrophotometer intensity was calibrated according to SOP 48.3 (attached in Appendix C) at the beginning of the experiment, in order to confirm that its operation was within the accepted criteria. The calibration protocol is summarised in Chapter 3. For the calibration of the wavelength and intensity accuracy similar tests were performed as described in Chapter 3. The target values and the measured values are shown in Appendix B for each one of the performed tests.

⇒ *Protein H*: The measured values from the wavelength accuracy tests (Appendix B, Table B1) were within the accepted limits for the region 187–586 nm. From the intensity performance tests (Appendix B, Table B2), the maximum of the ACS solution appeared to deviate by +0.8% from the target value at 291 nm, while the ratio of the signals at 193 nm and 291 nm was within the acceptance criteria and calculated as 2.08. The deviation for the (*R,R*)- and (*S,S*)-Na[Co^{III}(EDDS)] enantiomers was +3.8–4.6% at 211.5 nm.

⇒ *Somatropin (hGH)*: The wavelength accuracy criterion was met for the region 187–460 nm (Appendix B, Table B3), while at longer wavelengths (536–586 nm) a shift of −1.0 nm was observed. The intensity signal was within calibration at 291 nm (Appendix B, Table B4), while at lower wavelengths a deviation of +3.6% was observed for the ACS solution (at 193 nm) and +2.8% for the (*S,S*)-Na[Co^{III}(EDDS)] enantiomer (at 211.5 nm). The ratio of the signals at 193 nm and 291 nm for the ACS solution was calculated as 2.11, which is slightly higher than the accepted value (within 2.05–2.10).

⇒ *Protein A*: The wavelength accuracy criterion was met for the region 187–586 nm (Appendix B, Table B7). The intensity signal presented a deviation of −4.3% and −5.0% for the ACS solution at 193 nm and 291 nm respectively (Appendix B, Table B8), while the ratio of the two signals was calculated as 2.1 (within the accepted values). The intensity signal for the (*R,R*)-Na[Co^{III}(EDDS)] enantiomer (at 211.5 nm) presented a deviation of −4.2%.

⇒ *Etanercept (TNFR2:Fc)*: A wavelength shift of −0.9–1.6 nm was observed for the region 445–586 nm (Appendix B, Table B9). For the near-UV area (240–361 nm) the wavelength accuracy criterion was met, while for part of the far-UV

region (190–209 nm) a −1.0 nm shift was observed. The intensity signal presented a small deviation of +2.4% for the ACS solution at 291 nm (Appendix B, Table B10), and the ratio of the signals at 193 and 291 nm was calculated as 2.05 (within the accepted values). Intensity at 211.5 nm presented a deviation of +6.6 and +8.2% from the target value for the (*R,R*)- and (*S,S*)-Na[Co^{III}(EDDS)] enantiomers respectively.

4.2.7. Secondary structure analysis

SELCON3 and CDSSTR from CDPro (a suite of programs for protein CD analysis),[50] and SSNN were used for inferring the secondary structure of the tested proteins. For SELCON3 and CDSSTR the analysis was applied based on seven different reference sets (1–7), while for SSNN only reference set 7 enhanced by 5 more proteins was used (http://www2.warwick.ac.uk/fac/sci/chemistry/research/arodger/arodgergroup/research_intro/instrumentation/ssnn/). The reference sets contain CD spectra and secondary structures of soluble and denatured proteins, and differ in the number of proteins and wavelength range that they use in the analysis (http://sites.bmb.colostate.edu/sreeram/CDPro/).

For the studies of *Protein H, Somatropin (hGH), Protein A* and *Etanercept (TNFR2:Fc),* the analysed wavelength range was 180–260 nm, 185–240 nm or 190–240 nm, depending on each reference set's requirements as specified in the CDPro website. The secondary structure results are classified in different fractions according to each reference set. Reference sets 1, 3, 4, 6 and 7 use the same six classes: α-helix regular and distorted, β-sheet regular and distorted, turns and unordered; therefore, they are expected to give comparable results.[50,51] The regular and distorted fractions were summed up in order to give the helix and sheet content. Reference set 2 estimates α-helix, 3_{10}-helix, β-sheet, turns, poly(Pro)II structure and unordered.[52,53] In this case, the sum of α-helix and 3_{10}-helix resulted in the total helix content, while poly(Pro)II structure was added to unordered. For reference set 5 the same assignments apply as in reference set 2, but without the 3_{10}-helix fraction.[54]

4.3. PROTEIN H

The produced data for protein H are confidential and therefore have been excluded from the main thesis (Vol. 1). A separate Volume (Vol. 2) is provided for the examiners to review.

4.4. SOMATROPIN (RECOMBINANT HUMAN GROWTH HORMONE)

4.4.1. CD spectra of somatropin

CD spectra of somatropin (hGH) were recorded at the provided concentration (~3.56 mg/mL) using a demountable cuvette with a nominal path length of 0.001 cm, and after 2- and 4-fold dilution using a demountable cuvette nominally of 0.01 cm path length. Amino acid analysis, conducted by BioPharmaSpec, showed a concentration of 3.30 mg/mL for the original sample and therefore the values 3.30, 1.65 and 0.82 were used as the respective concentrations for conversion to Δε units, while the path lengths of the demountable cuvettes were assigned as described in materials and methods (Section 4.2.3). The processed data are shown in Figure 4.5 (left), with all three spectra presenting identical shapes, namely a maximum at 192.0 nm and two minima at 208.4 nm and 219.6 nm, typical of α-helical proteins and similar to previously reported studies of hGH.[55]

However, the spectra present different intensities which is unusual as Δε shows the per residue intensity and therefore is expected to be of similar levels for the same protein at different concentrations. Moreover, the theoretical negative maximum at 222 nm of a hypothetical protein with 100% α-helix content should not exceed −(12–13) $M^{-1}cm^{-1}$,[56-58] which is clearly not the case here, as hGH displays values of −14.2, −11.0 and −10.7 $M^{-1}cm^{-1}$ (minima at ~220 nm) for the initial, the 2- and 4-fold diluted samples respectively. Considering that the CD magnitudes were consistent among the three replicates in each case (see raw data in Appendix B, Figure B1), the above inaccuracy was a strong indication that an error was induced during conversion to Δε; from equation 1.3 it was concluded that this was caused by the assigned path length values. In order to confirm this suspicion, an additional spectrum was recorded following 30-fold dilution, using a 0.1 cm fixed path length cuvette (raw data in Appendix B, Figure B1). Indeed, the final processed spectrum (Figure 4.5, right) displayed a Δε value of approximately −6.25 $M^{-1}cm^{-1}$ at 208 nm, implying a 50–60% error in the previous calculations. The only explanation for such a large and consistent error in the assembly of the demountable cuvettes is that the high viscosity of the concentrated samples of hGH resulted in a path length increase, which was only obvious after converting the data to Δε units.

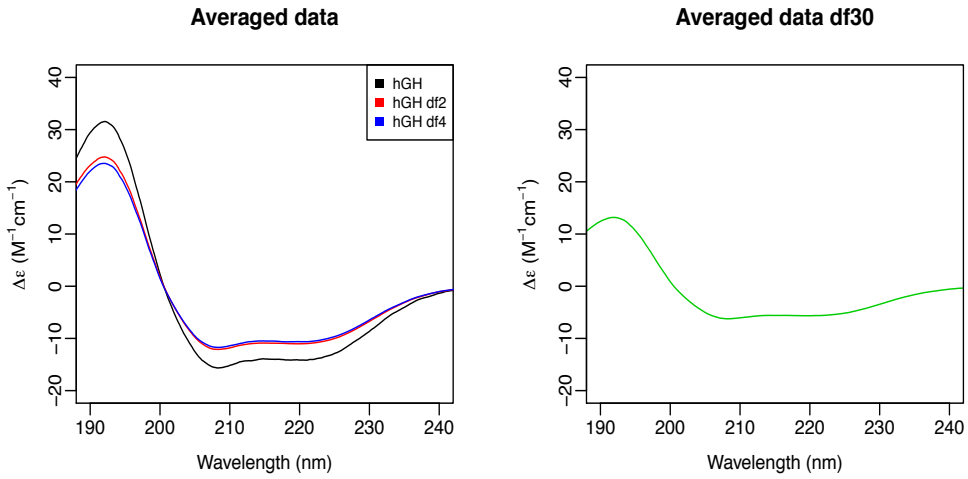


Figure 4.5: Normalised CD spectra (Δε units) of hGH at different concentrations averaged over three replicates. Left: hGH at its original concentration (nominally 3.56 mg/mL, black), after 2- (red) and 4-fold dilution (blue), using demountable cuvettes. Right: hGH after 30-fold dilution (green) using a fixed path length cuvette. For the conversion to Δε units, MRW 116.5 Da, path lengths of 0.000726 cm (for the original concentration), 0.00584 cm (for df2 and df4) and 0.1 cm (for df30) and concentrations of 3.30 mg/mL, 1.65 mg/mL, 0.82 mg/mL and 0.11 mg/mL were used respectively.

Given that the shape of the spectra does not change at different concentrations, indicating that the secondary structure of hGH is not affected after dilution, the path lengths of the demountable cuvettes were recalculated, by assuming equal Δε values for all four concentrations at 220 nm, based on the spectrum of the 30-fold diluted sample which presented a value of −5.68 $M^{-1}cm^{-1}$. This led to the following equation:

$$P_x = \frac{P \times \theta_x \times df_x}{\theta \times df} \qquad (4.1)$$

where P_x: is the path length of the demountable cuvette x (cm), θ_x: is the CD spectrum (mdeg) of the sample with a dilution factor df_x recorded in the demountable cuvette, θ: is the spectrum (mdeg) of the df-fold diluted sample recorded in the fixed cuvette with a path length of P (cm).

The average of the region 210–230 nm was used for the calculation and gave 0.001815 cm for the 0.001 cm cuvette and 0.01114 cm for the 0.01 cm cuvette, with the last being the average of the 2- and 4-fold diluted samples. The new values were used for conversion to Δε and the corrected spectra are shown in Figure 4.6. The CD spectra for all four samples look almost identical suggesting that the secondary structure of hGH does not change upon dilution. The slight variation at lower wavelength (192 nm) is perhaps due to the reduced resolution of the instrument at these wavelengths and increase of the noise level, as observed by the melting curves of

etanercept (Figure B10, Appendix B). The maximum, zero-crossing and minima points and values of the corrected normalised spectra (Δε units) are given in Appendix B, Table B5 for all three replicates of each sample.

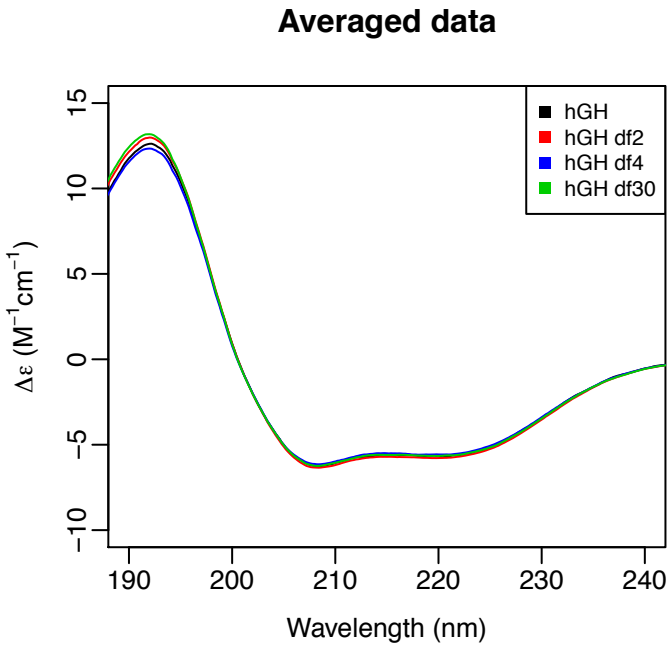


Figure 4.6: Corrected normalised CD spectra (Δε units) of hGH in different concentrations averaged over three replicates. Path lengths of 0.001815 cm (for the original concentration, black), 0.01114 cm (for df2 in red and df4 in blue) and 0.1 cm (for df30 in green) were used for conversion to Δε units.

Secondary structure estimation for somatropin

Similarly to the case of protein H, SELCON3, CDSSTR and SSNN were used in order to obtain secondary structure estimates of hGH from analysis of its CD spectra. The results from the 7 reference sets (for SELCON3 and CDSSTR) and 1 reference set for SSNN were compared and are presented with their NRMSD values in Appendix B, Table B6. The performed analyses with the different sets and programs were consistent for all four samples, and the average value and standard deviation with a 95% confidence interval of the 15 estimates were calculated for each sample and are shown in Table 4.6. The α-helix content of hGH at all four different concentrations was estimated as approximately 60% with a deviation of ±6%. This value is relatively close to previously reported estimates, which have been calculated from the hGH CD spectrum as 55±5%[16,19] and is also not very far from the X-Ray crystal structure which has shown that 100 amino acids in total are spread in four helical segments, resulting in 52% of α-helix.[18] The remaining 40% was spread among β-sheet conformation (~10%),

turns (~15%) and other (~15%), relatively close to the previously calculated with SELCON values of 2%, 12% and 20% respectively for a commercial formulation of hGH.[59] The random coil (or other) content presented the largest deviation of ±19% on average, perhaps due the difference of the CD spectra at lower wavelength (~192 nm)—the random coil spectrum normally presents a negative peak below 200 nm. This variation between the spectra at ~192 nm might also result from the higher noise levels of the instrument at lower wavelengths.

Table 4.6: Averaged values and standard deviations (with a 95% confidence interval) of the secondary structure estimates over the 15 analyses performed for hGH at different concentrations. SELCON3, CDSSTR (7 reference sets each) and SSNN were used for the CD spectra analysis.

	Helix	Sheet	Turns	Other
hGH	0.60±0.06	0.08±0.07	0.13±0.09	0.19±0.16
hGH df2	0.62±0.05	0.10±0.10	0.15±0.13	0.13±0.20
hGH df4	0.59±0.06	0.10±0.10	0.16±0.12	0.14±0.20
hGH df30	0.61±0.05	0.11±0.09	0.15±0.13	0.14±0.20

The original spectra and the calculated spectra from all analyses performed for hGH at the concentration provided and after 30-fold dilution are shown in Figure 4.7, while the data for the 2- and 4-fold diluted samples are shown in Appendix B, Figure B2. The spectral fits from all different programs and reference sets were very similar and in agreement with the original (experimentally measured) spectra, which is a good sign of an overall accurate estimation of the secondary structure of the protein. As expected and also seen in the previous study, SELCON3 presents slightly worse spectral fits, without this meaning that the actual results are of poorer quality.

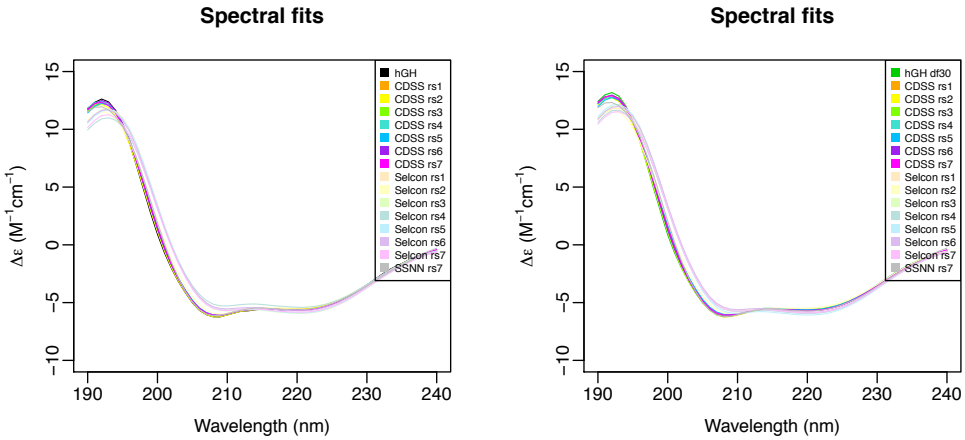


Figure 4.7: Original and calculated CD spectra of hGH at its original concentration (left) and after 30-fold dilution (right) from analysis with different software and reference sets. Spectral fits of the calculated spectra are in agreement with each other and with the experimentally measured spectra, suggesting an overall accurate secondary structure analysis.

In conclusion, correction of concentration errors—or path length errors, as was the case here—is very crucial, particularly in the case of proteins with high α-helix content, as it is generally accepted that α-helix estimates from most fitting programs are mostly affected by intensity errors. Cross-checking the secondary structure of hGH at different concentrations by analysis of its CD spectra (after these were corrected for the path length error) using different programs and reference sets, showed very close estimates and similar values in all cases, especially for the most important α-helix content. Therefore, the secondary structure of hGH is considered to be similar to previously reported studies and unaffected by changes in the concentration.

4.5. PROTEIN A

The produced data for protein A are confidential and therefore have been excluded from the main thesis (Vol. 1). A separate Volume (Vol. 2) is provided for the examiners to review.

4.6. ETANERCEPT

4.6.1. DLS spectroscopy of etanercept

Previous studies on the characterisation and stability of etanercept using DLS as one of the main techniques have shown hydrodynamic radii of 6.0–7.5 nm (or diameter of 12.0–15.0 nm) depending on concentration[60] and 7.0–8.3 nm (diameter of 14.0–16.6 nm),[61] while the theoretical value has been reported as 7.1 nm (diameter of 14.2 nm).[62] In this study, the hydrodynamic diameter of etanercept was measured as 14.1–15.1 nm (Z-average value as calculated by the Zetasizer software, based on intensity values),[63] which was in agreement with the values reported in the literature. However, this value was obtained only for the innovator samples (4–6), as they appeared to be nearly monodisperse (PDI < 0.1) presenting narrow distributions Figure 4.8 (left). The three biosimilar samples (1–3) presented wider peaks with a PDI of 0.4, indicative of moderate polydispersity, due to the presence of some larger particles of approximately 140 nm, which resulted in a Z-average value of ~20 nm. The distributions of all measurements and replicates for each sample are given in Appendix B (Figure B5) in intensity values in order to speculate for the presence of impurities, since intensity values are more sensitive to the presence of larger particles.

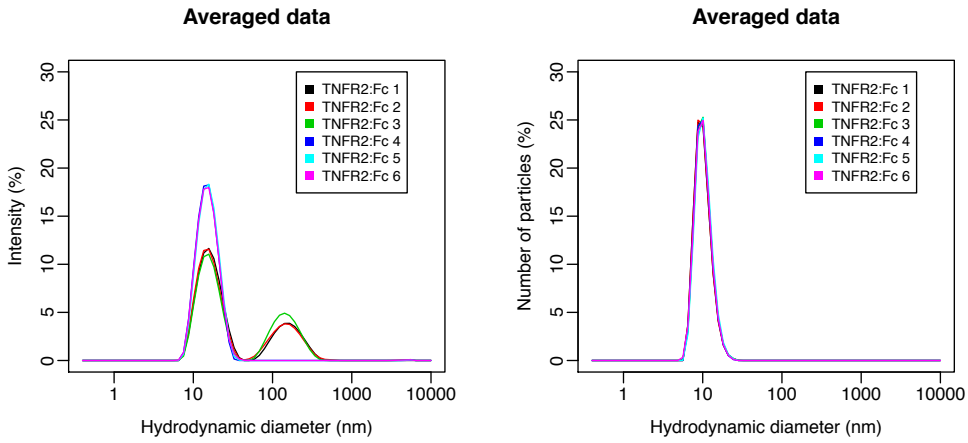
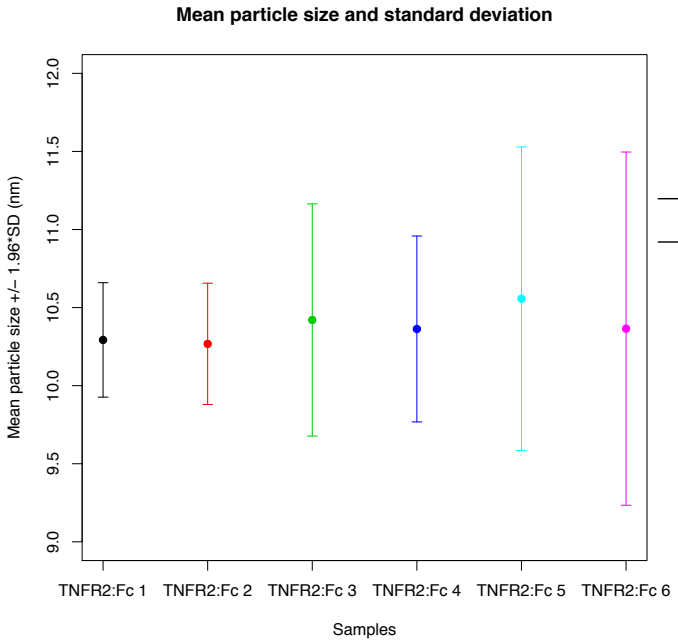


Figure 4.8: Left: Size distributions (nm) presented in intensity values for the six etanercept samples. The three innovator samples appeared to be of higher purity, presenting narrower peaks and an absence of the peak at ~140 nm, which was observed in the distributions of the biosimilar samples. Right: The size distributions in number values showed that the population of the 140 nm particles was insignificant. The presented data are averaged over 3 replicates (18 measurements in total).

The distributions of the samples in number values showed that the population of these large particles was not significant (Figure 4.8, right hand side panel), and all six samples presented identical peaks. The following mean sizes and the standard

deviations were calculated by the mean number values of each measurement (as calculated by the Zetasizer software), as these results are not affected by the presence of larger particles and represent better the most abundant species in the sample. For all six samples the overall mean diameters (over 18 measurements) were similar and calculated as 10.3–10.6 nm, while the standard deviations differed, with samples 3, 5 and 6 presenting more widely spread sizes (Figure 4.9), perhaps due to the relatively short equilibration time (60 s) that was allowed before the measurements started. The DLS of the biosimilars does suggest the presence of some large particles.



Sample	Mean ± 1.96*SD
TNFR2:Fc 1	10.3±0.4
TNFR2:Fc 2	10.3±0.4
TNFR2:Fc 3	10.4±0.7
TNFR2:Fc 4	10.4±0.6
TNFR2:Fc 5	10.6±1.0
TNFR2:Fc 6	10.4±1.1

Figure 4.9: Left: Overall mean particle diameters (nm) of etanercept indicated as solid circles. For each sample, the overall mean was calculated over 18 measurements using the number mean values as estimated by the Zetasizer software. The error bars indicate the standard deviation with a 95% confidence interval. The values of the overall means and the standard deviations for each sample are shown in the attached table (right).

Temperature experiment

The six etanercept samples were heated gradually in a temperature range from 25 to 75 °C and the intensity distributions throughout and at the end of the temperature experiments, when the samples were cooled down to 20 °C, are shown in Appendix B, Figure B6. In this case all samples presented impurities of larger sizes than the monomeric protein, however the number distributions confirmed that the population of these particles was small (inset). Figure B7 in Appendix B shows the size distributions at the beginning (at 25 °C, left) and at the end of the temperature experiment (cooled to 20 °C, right). All samples presented wider peaks and a shift to larger sizes than the

original measurement when cooled back down to 20 °C, suggesting the appearance of larger particles, which affected the shape of the distribution. These particles (when cooling down to 20 °C), were likely to represent partially unfolded etanercept molecules, while loss in intensity was probably a sign of a mixture of particles with different sizes.

Figure 4.10 (left) shows the mean size of the particles (number values averaged over six measurements) observed in each sample from 25–75 °C and when the samples were cooled down to 20°C. All six samples seemed to follow a common path throughout the experiment, showing initially a decrease to a size of 6.7–7.7 nm at 70 °C, while a slight increase in size was observed at 75 °C (7.0–8.4 nm) and a "jump" towards larger values (11.0–11.9 nm) after the samples were cooled back down to 20 °C. The observed size decrease represents a more compact conformation of the protein, which probably resulted from gradual loss of the interactions between the protein and arginine HCl—which was one of the excipients in the provided samples for the repeat analysis for the fluorescence data (2^{nd} analysis, Section 4.2.1)—as the temperature increased. A study by Arakawa *et al.* on the effect of arginine HCl on suppression of protein interactions has shown that the molecule acts by increasing the surface tension of water, and interacts with most amino acid side chains and peptide bonds, but presents limited binding to the outer surface of the protein; therefore, with a diameter of 0.76 nm, arginine HCl is expected to cause a significant expansion in the size of the protein.[64] Another study has shown that the concentration-dependence of the hydrodynamic diameter of etanercept is suppressed by addition of arginine HCl in the protein solution, resulting in a size of 13.0–13.7 nm (Z-average value), namely a size increase in the order of 16% and 37% for high (50 mg/mL) and low (5 mg/mL) concentration samples respectively.[65] These findings are in agreement with the size decrease that was observed in the present study, which was in the order of 40% at 70 °C. Furthermore, the three innovator samples appeared slightly larger than the biosimilar samples consistently throughout the temperature experiment, which was also attributed to arginine. Since the innovator samples were subjected to 2-fold less dilution than the biosimilar samples—the first were provided at a concentration of 25 mg/mL while the second at 50 mg/mL, and all samples were diluted to 1 mg/mL—they contained twice as much arginine HCl, which probably resulted in a somewhat larger size. Interestingly, all six samples presented very similar sizes at 55 °C (8.1–8.4 nm), which seems to be a key temperature. The increase in size at 75 °C was probably due to partial unfolding of the structure, while the "jump" at the end of the experiment

(cooled to 20 °C) showed size values relatively close to the initial ones but of larger particles, perhaps due to combination of two events: restoration of the interactions between arginine and etanercept and partial unfolding of the protein. The standard deviations—calculated over the averaged number values of six measurements and shown as error bars in Figure 4.10 (right)—were smaller compared to the above-discussed measurements at room temperature, perhaps due to the longer equilibration time (180 s) that was allowed before each set of measurements. Larger deviations were evident at 75 °C and at the end of the temperature experiment (cooled to 20 °C), indicating a mixture of protein particles with varying sizes.

A limitation of the DLS instrumentation is that the temperature maximum is 75 °C, which is a critical point for the case of etanercept, as according to previously published studies it presents 3 transition melting temperatures due to independent unfolding of the different domains (TNFR and C_H2, C_H3 of the Fc component). While TNFR and C_H2 are expected to unfold at 57.0 °C and 69.6 °C respectively, the C_H3 domain presents a higher melting temperature, at 82.8 °C.[37,60] Consequently, complete unfolding of the protein was precluded in the DLS temperature experiment, resulting in mixed particle sizes and large standard deviations at its closure. The average sizes and the standard deviations at each temperature can be found in Appendix B (Table B11).

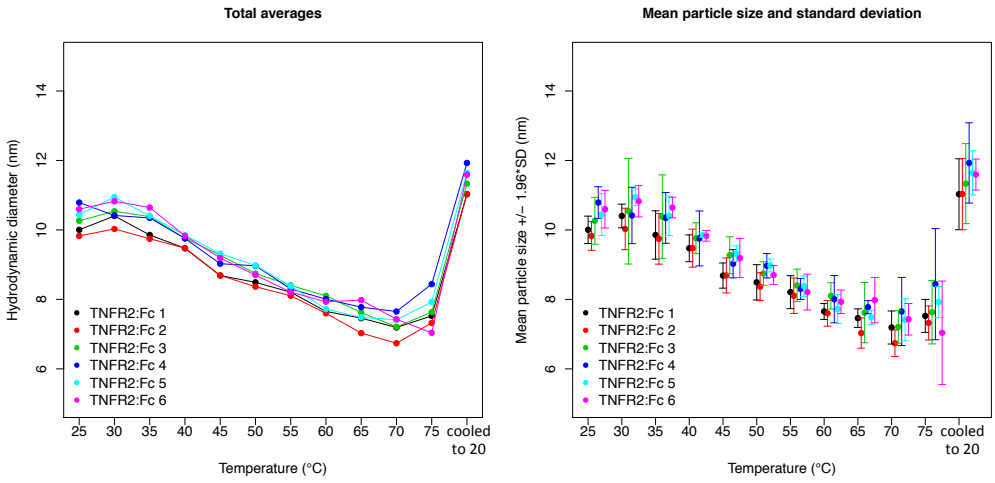


Figure 4.10: Hydrodynamic diameter (nm) of the six etanercept samples over a range of temperatures from 25–75 °C and after the samples were cooled back down to 20°C. Left: the averaged sizes over 6 measurements are shown. Right: mean particle diameters (solid circles) and standard deviation with a 95% confidence interval (error bars). The results are presented in terms of number values.

In conclusion, the DLS data showed similarity in size and thermal stability among the biosimilar and the innovator samples. The innovator samples, however, appeared to be of higher monomeric purity compared to the biosimilars, which

presented larger particles with a size of ~140 nm—affecting the intensity distributions—and wider peaks. The temperature experiment showed common trends in size changes for all 6 samples and expansion of the structure at the end of the experiment, after cooling down to 20 °C. The concomitant widening of the peaks indicated the presence of particles with mixed sizes, which probably indicates that unfolding was not fully completed. No prominent aggregation was evident for any of the samples, perhaps due to the relatively low concentration of the protein and the protective effect of arginine HCl against aggregation.[60,65]

4.6.2. CD spectra of etanercept

Far-UV

CD spectra of etanercept were recorded at the provided concentrations (~0.9 mg/mL) using a demountable cuvette with a nominal path length of 0.01 cm. Raw data of the buffer and samples are shown in Appendix B, Figure B8 (replicates and averages). The processed data (converted to Δε units) are shown in Figure 4.11 (left). All samples presented identical shapes, namely a negative maximum at around 193 nm and a second negative peak of lower intensity at 230 nm, similarly to previously reported far-UV CD spectra of etanercept.[60,66] This kind of CD spectrum is unusual, dominated by random coil structure and β-sheet conformation perhaps of type II (as described for protein A, Section 4.5.1 in Vol. 2). The variation in intensity of the six samples probably results from a concentration error or can be partly attributed to the observed variation among replicates in the raw data (see for example replicates for samples 3 and 6, Appendix B, Figure B8). The minima and the Δε values for each sample (averaged over three replicates) are given in Appendix B (Table B12), and these spectra were used for the secondary structure analysis as discussed below. In order to eliminate the intensity variation due to concentration error and to aid visual inspection of the shape of the spectra, the data were further normalised based on the minimum Δε value of sample 1 (Figure 4.11, right). Similar shape for all 6 samples was evident, indicating similarity in their secondary structure and implying that the intensity variation was induced during conversion to Δε, both due to inaccuracy in concentration estimation and human error in the assembly of the demountable cuvette.

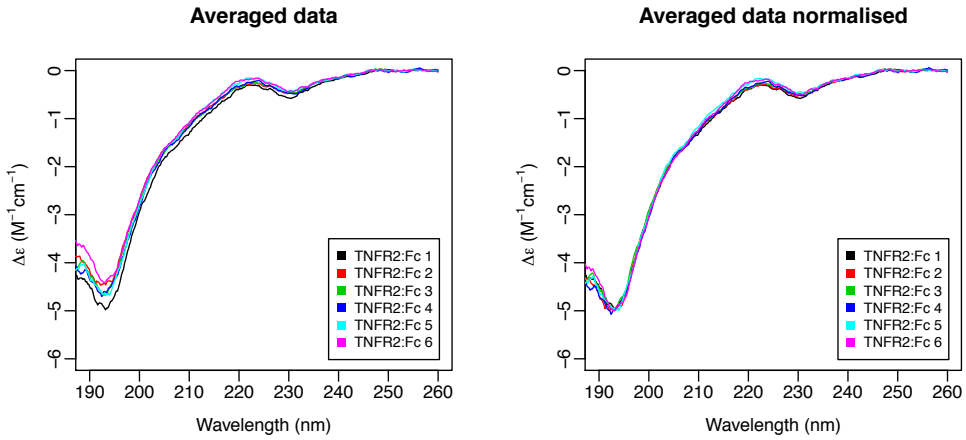


Figure 4.11: CD spectra (Δε units) of the six etanercept samples at their original concentration (~0.9 mg/mL) in ammonium bicarbonate buffer using a 0.01 cm demountable cuvette (left). Right: The spectra were normalised to the intensity of the negative maximum at 193 nm of sample 1, in order to aid visual inspection of their shape. Data were acquired at room temperature and are presented as an average over 3 replicates (12 scans each).

Melting curves

In order to inspect the effect of temperature in the CD spectrum of etanercept, sample 1 was recorded as an example in a range from 20 °C to 100 °C and after cooling the sample down to 20 °C (Appendix B, Figure B9). The onset of the protein unfolding was observed at 50 °C, while the structure was fully unfolded at 80 °C with the negative maximum presenting a shift from 193 nm to 202 nm, suggesting an increase in the random coil content. Upon cooling, the peak shifted slightly to shorter wavelengths (~199 nm), accompanied by a change in its shape perhaps suggesting partial refolding of the protein. Refolding of etanercept has been reported to be favoured by arginine HCl,[65] which was one of the excipients in the samples from the 2nd analysis used for the melting curves. Furthermore, the CD spectra cross in more than one point, at around 195 nm (isodichroic points), indicating that etanercept undergoes more than one transitions.[67]

The signals of the crossing-point at 195.4 nm and of the ending (at high temperatures, ≥80 °C) negative maximum at 202 nm were followed for all 6 samples throughout a range of temperatures from 25 °C to 90 °C resulting in the melting curves shown in Figure 4.12 (left and right respectively). The two wavelengths presented similar transitions for all six samples, indicating that the biosimilar products are of equal thermal stability to the innovator ones. The intensity of the signal at 195.4 nm was almost stable up to ~57 °C, when it started increasing, while at 75 °C it presented a decrease towards more negative values. The signal at 202 nm started decreasing at

~43 °C and was constant from ~82 °C onwards. As discussed above (in the DLS results, Section 4.6.1), etanercept presents three melting temperatures, which vary depending on the concentration of the sample. The results from two published studies that have used differential scanning calorimetry (DSC) varied significantly, showing melting temperatures at: T_m1 = 51.2/57.0 °C, T_m2 = 60.2/69.6 °C and T_m3 = 85.5/82.8 °C for the first/second study respectively, at a protein concentration of 1 mg/mL, while at 0.25 mg/mL etanercept was less stable with T_m1 = 42.9 °C, T_m2 = 57.0 °C and T_m3 = 85.6 °C.[37,60] In the present study, however, a tentative attempt was made to assign the melting temperatures from the melting curves, and these were found approximately as: T_m1 = 53 °C, T_m2 = 70 °C (as the average of 68 °C and 72 °C) and T_m3 = 82 °C. The raw melting curves, with data points every 0.1 °C, are shown in Figure B10 (top, Appendix B), illustrating the levels of noise from the instrument and the deterioration of the resolution especially at lower wavelength. The respective HT spectra are also included (Figure B10, bottom) showing that aggregation might have occurred for some of the samples (*e.g.* samples 2, 3, 4 and 6), however to a very limited extent due to the minor changes in the HT values. No solid particles were observed in any of the samples at the end of the experiment.

CD spectra and the respective HT spectra of the six samples recorded at the beginning (25 °C) and at the end of the melting curves (cooled to 25 °C) are presented in Appendix B, Figure B11 and Figure B12. For all six samples, the HT spectra at 25 °C and after the samples were cooled down to 25 °C are identical (Figure B12), suggesting that no significant aggregation took place. All six samples present CD spectra of similar shape both at the beginning and at the end of the experiment (Figure B11), but samples 2 and 5 differ somewhat in intensity at the end of the experiment, perhaps due to buffer evaporation. The spectra at 25 °C suggest a Δε value of approximately −2.4 for the minimum at 192 nm, revealing a +50% error in the calculated Δε spectra for the above-described high concentrated samples (~0.9 mg/mL) at room temperature, which presented a value of approximately −4.8 (Figure 4.11, left). As shown in the preceding studies, this error results from the path length assignment of the demountable cuvette and seems to be quite common in cases of concentrated samples, where increased viscosity affects assembly of the cuvette. More specifically, high viscosity seems to be related to the absence of excipients such as surfactants (*e.g.* Tween 20), amino acids (*e.g.* histidine or arginine), and carbohydrates (*e.g.* sucrose or trehalose), which are usually added to biopharmaceutical formulations in order to reduce viscosity and increase stability of the preparation.[43,44,65] Absence of excipients

was the case for all proteins presented so far in this chapter and also for etanercept, as its CD spectra were collected in ammonium bicarbonate buffer (samples provided for the 1st analysis for which spectra were recorded at room temperature), while the path length error was not observed for the trastuzumab samples in Chapter 3, as these were properly formulated, containing Tween 20, histidine and trehalose (see Figure 3.27). However, the purpose of this study was comparison of the similarity between biosimilar and innovator samples, rather than estimation of the actual secondary structure contents. Therefore, the Δε spectra in this case were used without applying any correction of the path length.

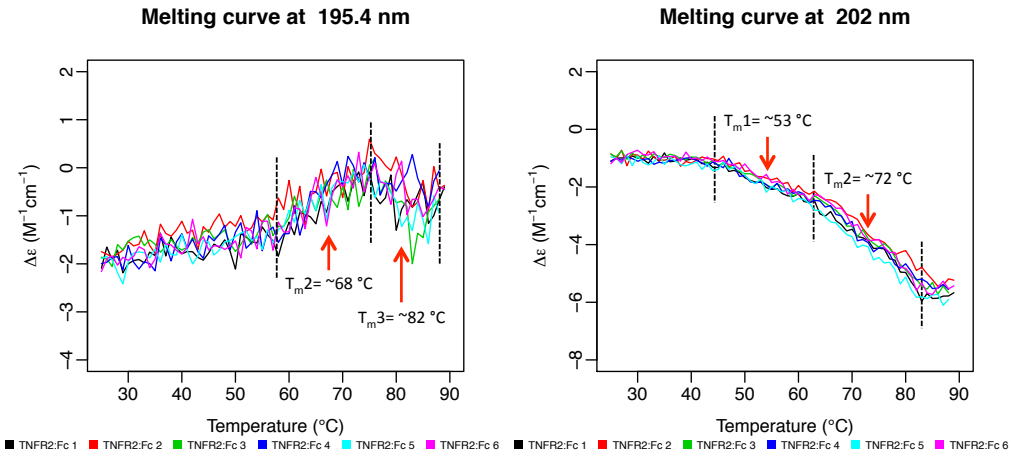


Figure 4.12: Melting curves (Δε units) of the etanercept samples at one of the crossing points of the spectra at 195.4 nm (left) and at the ending negative maximum at 202.0 nm (right). The curves were recorded for the formulated samples that were diluted in two steps: to a concentration of 1 mg/mL with PBS, and a concentration of 0.1 mg/mL with water. The 0.1 cm fixed path length cuvette was used.

Near-UV

Near-UV CD spectra of etanercept were recorded at the provided concentrations (~0.9 mg/mL) using a 1 cm cuvette. The raw data of the samples and the buffer are shown in Appendix B, Figure B13 (replicates and averaged data), while the processed spectra are presented in Figure 4.13. All samples present CD spectra of similar shape, while sample 1 presents a slightly higher intensity and a somewhat different shape especially in the tyrosine region (peak at 275 nm). The well-defined CD spectrum in the near-UV region is similar to previously published data[66] and confirms that etanercept presents ordered structure. All six samples appear to have similar tertiary structure within experimental error, with sample 1 being slightly different from the rest in the environment of its tyrosine residues.

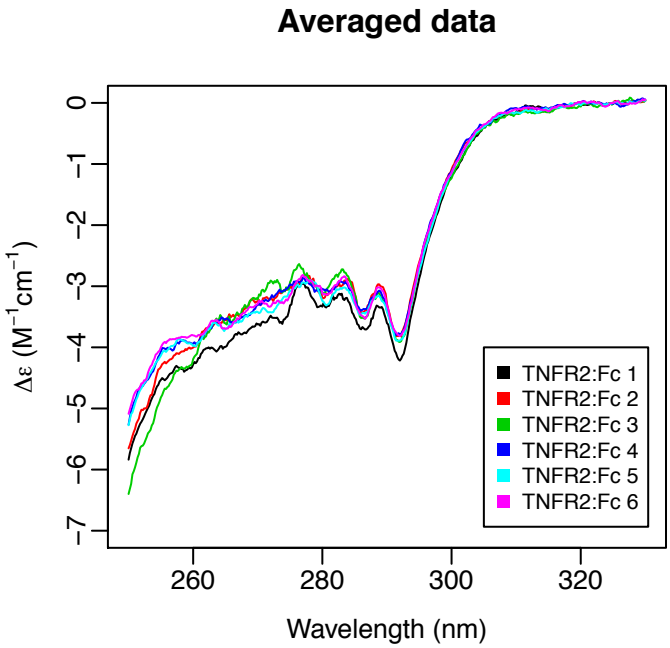


Figure 4.13: Near-UV CD data (Δε units) of the six etanercept samples in their original concentrations (~0.9 mg/mL) using a 1 cm cuvette.

Secondary structure estimation for etanercept

In order to have a "reference" secondary structure of the protein for comparison with the obtained results from the analysis with CDSSTR, SELCON3 and SSNN (following), the secondary structure of etanercept was calculated by combination of the PDB crystal structures for TNFR2 and the Fc fragment of human IgG1, as described above (Section 4.1.4), and gave the following results:

Table 4.7: Secondary structure percentages of etanercept as calculated by the PDB entries for TNFR2 and the Fc fragment of human IgG1 (as discussed in introduction for etanercept, Section 4.1.4). According to [68], for the α-helix content, 3_{10}-helix was added to α-helix; β-strand gave β-sheet; turns content was the sum of turns and bends; β-bridge conformation and residues that were not assigned any secondary structure were summed to give other.

α-helix = H+G	β-sheet = E	Turns = T+S	Other = B+NA
4.9%	33.4%	17.2%	44.5%

Similarly to what was applied to protein A (secondary structure estimation, Section 4.5.1 in Vol. 2), 3_{10}-helix (G) was added to α-helix (H), β-strand (E) gave the β-sheet content, bend (S) was added to turn (T), and the sum of β-bridge (B) and not assigned (NA) structure gave other.

The CD spectra of the etanercept samples were analysed for their secondary structure content using CDSSTR and SELCON3 with reference sets 1–7, and SSNN, which uses reference set 7. The results from all performed analyses are presented in Table B13 (Appendix B), where a large variation was apparent among estimates from different software and reference sets, especially for the random coil content. The analyses that gave results close to the calculated values from the PDB crystal structures (≤10% deviation) are presented in Table 4.8. CDSSTR with reference sets 3 and 4, and SELCON3 with reference set 4 fulfilled the structure criterion, with CDSSTR showing similarity for all 6 samples, namely 2–3% α-helix, 33–37% β-sheet, 24–27% turn and 35–39% other, while SELCON3 showed some variation among samples estimating a higher α-helix (~10%) and other content (35–47%) and lower sheet (~27%) and turns (~22%).

Table 4.8: Results from the secondary structure analysis of the CD spectra of etanercept with different software and reference sets. Only the programs/reference sets that gave results close (deviation ≤10%) to the calculated ones from the crystal structures are shown. CDSSTR rs3, 4 and SELCON3 rs4 fulfilled the structure criterion.

Program	Sample	Helix	Sheet	Turns	Other
CDSSTR rs3	TNFR2:Fc 1	0.03	0.37	0.24	0.37
	TNFR2:Fc 2	0.02	0.35	0.24	0.38
	TNFR2:Fc 3	0.02	0.33	0.26	0.39
	TNFR2:Fc 4	0.02	0.37	0.24	0.37
	TNFR2:Fc 5	0.02	0.36	0.24	0.38
	TNFR2:Fc 6	0.02	0.34	0.25	0.39
CDSSTR rs4	TNFR2:Fc 1	0.03	0.35	0.27	0.35
	TNFR2:Fc 2	0.03	0.34	0.26	0.37
	TNFR2:Fc 3	0.03	0.33	0.26	0.39
	TNFR2:Fc 4	0.03	0.35	0.27	0.36
	TNFR2:Fc 5	0.02	0.35	0.26	0.37
	TNFR2:Fc 6	0.03	0.33	0.26	0.38
SELCON3 rs4	TNFR2:Fc 1	0.02	0.27	0.25	0.47
	TNFR2:Fc 2	0.11	0.33	0.21	0.35
	TNFR2:Fc 3	0.10	0.27	0.22	0.41
	TNFR2:Fc 4	0.10	0.24	0.20	0.46
	TNFR2:Fc 5	0.10	0.27	0.22	0.41
	TNFR2:Fc 6	0.12	0.28	0.21	0.39

Table 4.9 shows the overall means and standard deviations for each structure, from the analyses with all software and reference sets. All six samples presented similar secondary structure contents consisting of: 2–3% α-helix, 21–23% β-sheet, 20–21% turn and 54–57% other. Deviations with a 95% confidence interval were

relatively small for the helix estimates (±6%), sheet and turns showed a deviation of ±18% and ±14% on average respectively, while other content presented the largest deviation of ±27% on average.

Table 4.9: Overall means and standard deviations with a 95% confidence interval from the analyses with all different software programs and reference sets for each etanercept sample. The results are the average of 15 analyses.

	Helix	Sheet	Turns	Other
TNFR2:Fc 1	0.02±0.05	0.21±0.17	0.20±0.16	0.57±0.30
TNFR2:Fc 2	0.02±0.07	0.23±0.16	0.20±0.13	0.55±0.25
TNFR2:Fc 3	0.02±0.06	0.22±0.20	0.21±0.15	0.54±0.28
TNFR2:Fc 4	0.02±0.06	0.23±0.21	0.21±0.15	0.54±0.29
TNFR2:Fc 5	0.02±0.06	0.23±0.20	0.21±0.14	0.54±0.28
TNFR2:Fc 6	0.03±0.07	0.21±0.16	0.20±0.13	0.57±0.23

The original CD spectra of the etanercept samples and the calculated spectra, after the analysis with programs and reference sets that gave results close to the crystal structure estimates for at least one of the samples, are shown in Figure 4.14. Fits for CDSSTR rs2–4 and SELCON3 rs2–4 are presented, and as can be observed fits of SELCON3 rs4 were of the poorest quality in most of the cases, SELCON3 rs3 estimated the second feature of the etanercept CD spectrum (at 230 nm), but shifted at around 215–220 nm, while the fits from SELCON3 rs2 varied from poor (for samples 1, 2 and 6) to acceptable quality (for samples 3, 4 and 5). CDSSTR presented improved spectral fits, which is generally expected as described for protein H in Vol. 2 (see secondary structure estimation for protein H Section 4.3.1).

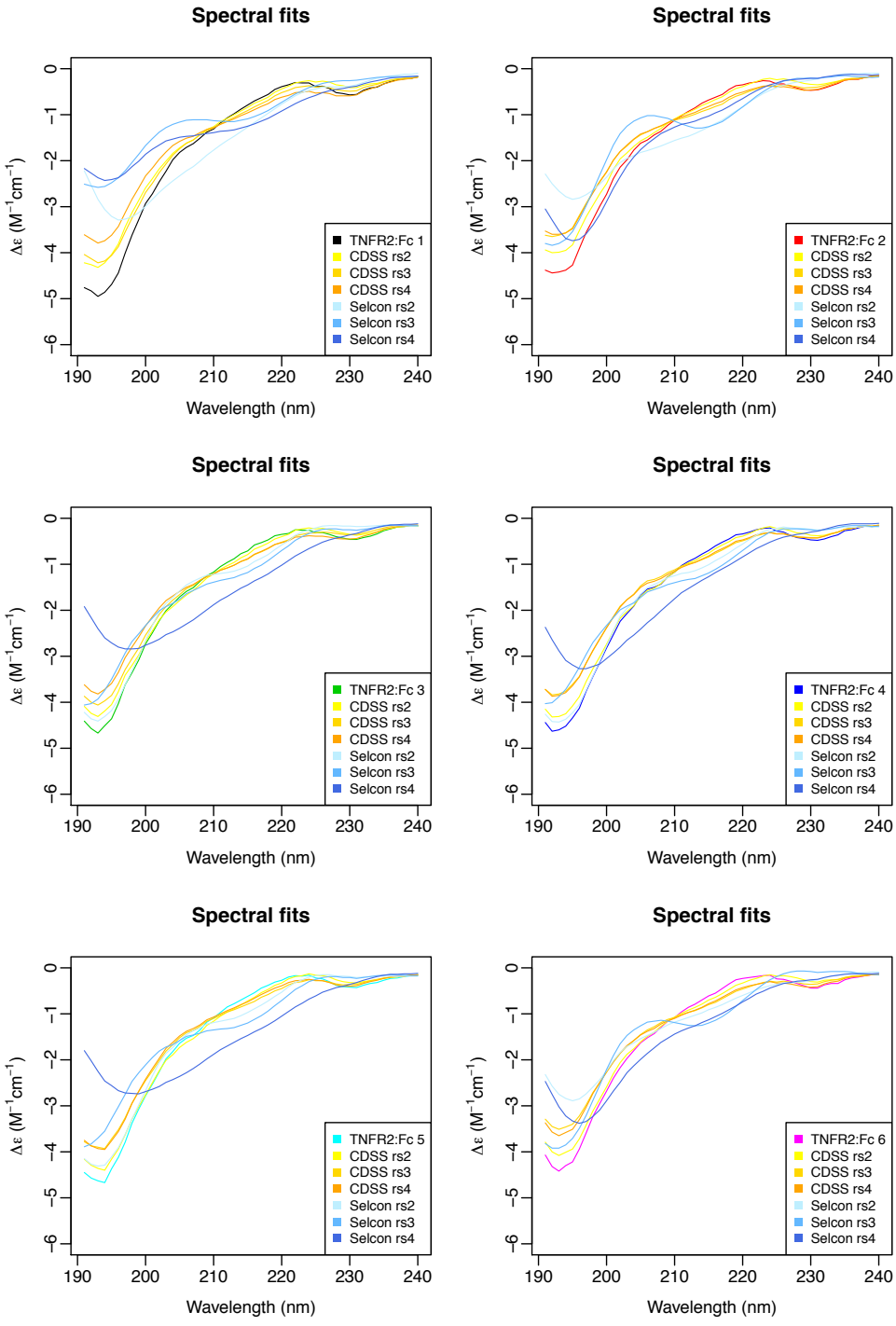


Figure 4.14: Original (experimentally measured) and fitted CD spectra of the six etanercept samples after analysis with different software and reference sets. Spectral fits of programs that presented ≤10% deviation from the PDB-calculated secondary structure for at least one of the samples are shown.

In conclusion, the far-UV CD spectra for all six samples of etanercept look identical in shape, dominated by random coil and β-sheet conformation, suggesting similarity in their secondary structure. The average from analyses of the CD spectra

with various software and reference sets confirmed similarity in secondary structure contents among samples, however the results are probably not representative of the actual structure of etanercept due to the large variation among estimates from different software and reference sets. Melting curves recorded at two wavelengths presented similar transitions for all six samples, indicating that the samples were of equal thermal stability. Onset of protein unfolding was observed at around 50 °C, the structure seemed to reach its fully unfolded conformation at approximately 80 °C, while all six samples displayed partial reversibility of unfolding and similar spectra at the end of the temperature experiment when cooled down to 20 °C. The near-UV CD spectra appeared to be of similar shape, suggesting that all six samples had similar tertiary structure with sample 1 being slightly different from the rest.

4.6.3. FT-IR spectroscopy of etanercept

FT-IR spectra of etanercept were recorded at the provided concentrations (~0.9 mg/mL) in ammonium bicarbonate buffer (1st analysis). The processed data (corrected for the buffer/water vapour and zeroed) for all replicates are shown in Figure B14 (Appendix B). The apparent replicate-to-replicate variation, more obvious for samples 1 and 6, was probably induced during buffer correction, while raw data (not shown) appeared more consistent (but dominated by liquid water). The buffer was recorded once, in the beginning of the measurements, therefore it is likely that small changes in humidity during the rest of the measurements have affected correction of the spectra and were reflected in the processed data, while the small concentration of the samples (10 mg/mL is deemed to be appropriate for FT-IR) contributed to exaggeration of these variations. The averaged data (over three replicates) for the six samples are shown in Figure 4.15 (left). The features of the spectra are almost identical for the six samples, presenting the amide I band at around 1638 cm^{-1} and the amide II at ~1548 cm^{-1}. The amide I band was normalised to unity for all samples (Figure 4.15, bottom right), and was subsequently resolved with the second derivative (top right) in order to aid peak positioning. The second derivative revealed the main feature of the etanercept spectrum at 1638 cm^{-1}, which in combination with the peak at 1685 cm^{-1} is characteristic of antiparallel β-sheet structure, indicating that the main conformation of the protein is indeed β-sheet. Additional second derivative peaks appeared at 1654 cm^{-1}, denoting presence of α-helix or random coil, at 1662 cm^{-1} and 1672 cm^{-1}, which are typical of turns and loops.[69,70] For all six samples the normalised amide I band was input to the IR-reference set retrained version of the SSNN software (as described in Chapter 2) in order to provide estimates of the secondary structure contents.

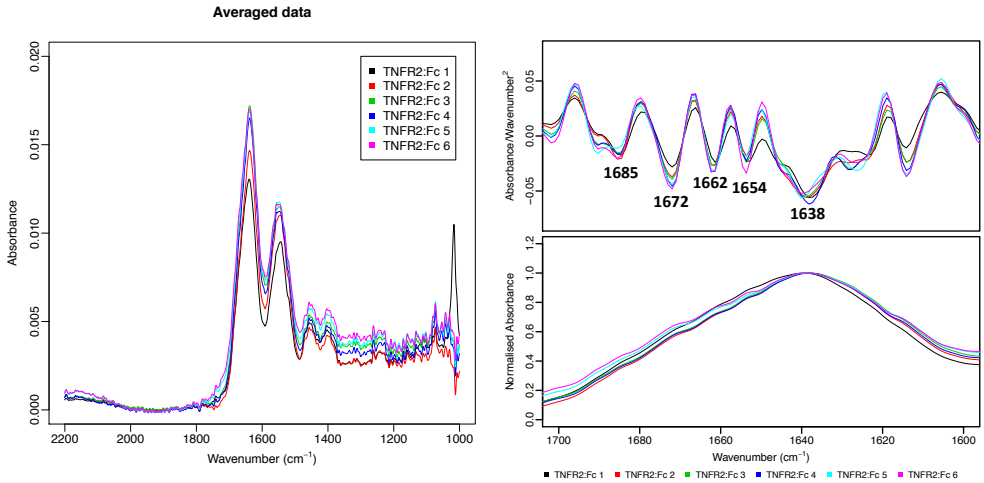


Figure 4.15: Left: ATR-FTIR spectra of the six etanercept samples (~0.9 mg/mL) after buffer subtraction, water vapour correction and zeroing. The spectra are presented as an average over three replicates in absolute absorbance units. Right: The amide I band was normalised to unity (bottom). Second derivative amide I spectra (top) were used to aid peak positioning.

Secondary structure analysis for etanercept

Table 4.10 shows the results from the secondary structure analysis of the six etanercept samples and the NRMSD values, which indicate a reasonable fit between the measured and the modelled spectrum. The estimates indicate that there are no significant differences between the secondary structures of the samples. Sheet content was estimated as approximately 33%, which was close to the value that was calculated from the PDB crystal structures and also to the CDSSTR rs3 and 4 results from the CD spectra analysis. Turn was estimated as 24% (sum of turn and bend), which was similar to the CDSSTR (rs3, 4) estimates, while helix was overestimated (20%) with a consequent underestimation of the random coil content (22%). This confusion in FT-IR spectra analysis is due to the fact that the average maxima for both α-helix and random coil bands are at 1654 cm^{-1}, with their extremes differing only slightly: 1648–1657 cm^{-1} and 1642–1657 cm^{-1} respectively.[71] Therefore differentiation between the two structures is challenging in most of the cases. The small deviations that appeared in samples 1 and 6 (~5% and ~3% in sheet content respectively) were probably the result of the variation in their replicates.

In conclusion, the FT-IR spectra of all six etanercept samples appeared similar within experimental and processing error. As discussed in Chapter 2, the results from the quantitative analysis of the spectra with the use of SSNN should be regarded as an approximation, as overlap of different bands (such as α-helix and random coil) can confuse the algorithm and disrupt the estimation.

Table 4.10: Secondary structure assignment for the FT-IR spectra of etanercept using the IR trained version of SSNN.

	Helix	Sheet	Bend	Turn	Other	NRMSD
TNFR2:Fc 1	0.24	0.28	0.11	0.14	0.23	0.024
TNFR2:Fc 2	0.20	0.34	0.11	0.13	0.22	0.030
TNFR2:Fc 3	0.20	0.33	0.11	0.13	0.22	0.037
TNFR2:Fc 4	0.20	0.34	0.11	0.13	0.22	0.032
TNFR2:Fc 5	0.20	0.33	0.11	0.13	0.22	0.050
TNFR2:Fc 6	0.23	0.30	0.11	0.14	0.23	0.055

4.6.4. Fluorescence spectroscopy of etanercept

Fluorescence emission spectra for the etanercept samples were recorded for two excitation wavelengths: 280 nm (where tyrosine and tryptophan residues fluoresce) and 295 nm (only tryptophans fluoresce). Two analyses were conducted in different buffers.

For the **1st analysis**, the samples were provided as buffer exchanged into ammonium bicarbonate buffer and were recorded as such at the provided concentrations (~0.9 mg/mL). Raw fluorescence data (in three replicates) for the samples and the buffer are shown in Appendix B, Figure B15–Figure B16 for both excitation wavelengths (280 nm and 295 nm). The data following subtraction of the ammonium bicarbonate signal are shown in Figure 4.16 (top). A prominent difference between the emission spectra of the biosimilar and the innovator samples was the observed additional peak at around 460 nm for the biosimilar samples, perhaps suggesting the presence of an aromatic compound. All six samples presented their maxima at ~334.8 nm and ~336.5 nm for excitation wavelengths 280 nm and 295 nm respectively (Table 4.11). The similarity of the emission wavelengths for 280 nm and 295 nm excitation indicates dominance of tryptophan fluorescence with both excitation wavelengths. The observed differences in intensity among samples suggest either differences in the environment of the aromatic (Tyr/Trp) residues or differences in concentration. A decrease in the fluorescence intensity is usually indicative of Tyr/Trp residues that are exposed to the solvent.[45] However, significant changes in the environment such as exposure to the more polar environment of the solvent would have also caused a shift in the maximum wavelength towards higher wavelengths (closer to the maximum for Trp in water, at 353 nm),[46] which was not observed in this case for any of the samples. The intensity of fluorescence of Trp and Tyr residues can also be affected by complicated quenching mechanisms that arise from neighbouring amino acids in the folded state of the protein, causing a signal decrease.[72] It is likely

therefore, that either such quenching mechanisms or a mis-calculation of the concentrations was responsible for the intensity variation.

The fluorescence intensities were divided by the molar protein concentration to account for variation in the protein concentration between samples. These data are shown in Figure B17 (Appendix B). Sample 1 presented much higher intensity than the rest, from which it can probably be inferred that either its concentration was underestimated or sample 1 indeed differed in the aromatic residues environment. Referring back to the near-UV CD spectra of the six samples (Figure 4.13), sample 1 presented somewhat different shape in the near-UV wavelength (250–280 nm), in the tyrosine area. Moreover, by looking at the far-UV CD spectra (Figure 4.11, left) a higher intensity for sample 1 was apparent. Consequently, the variation in the fluorescence intensity of sample 1 probably results from the combination of both factors.

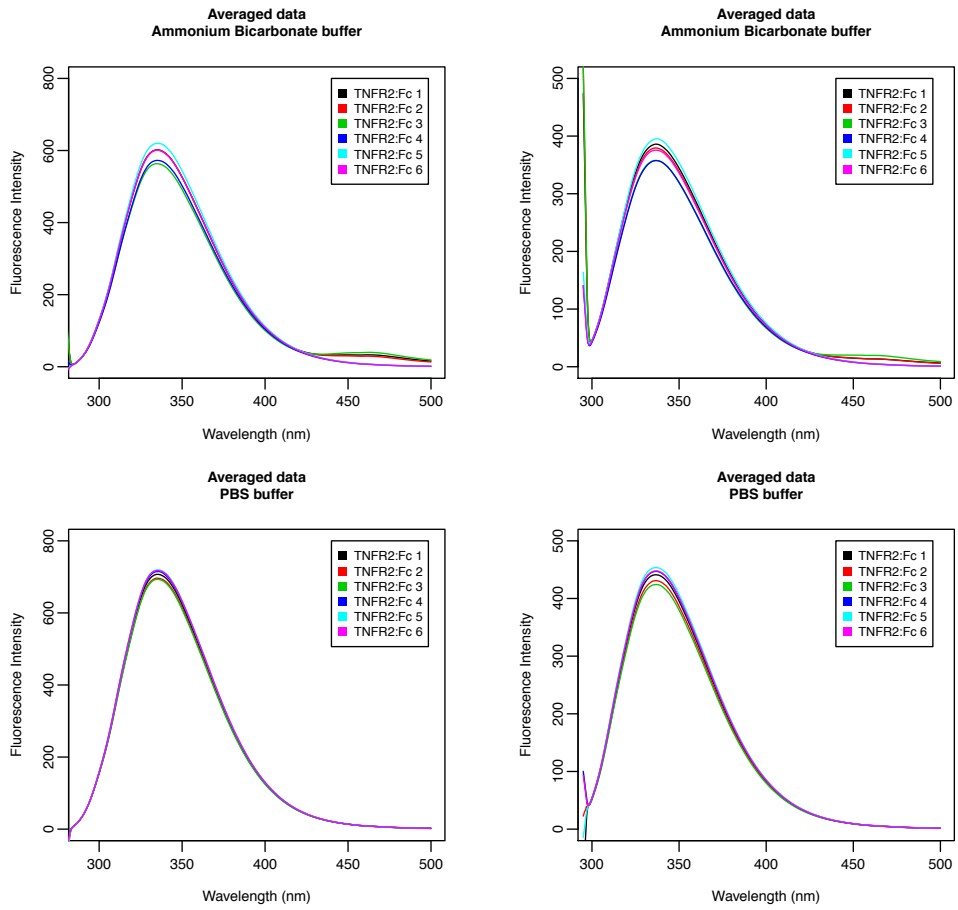


Figure 4.16: Baseline corrected and averaged (over three replicates) fluorescence emission spectra of the etanercept samples for excitation wavelengths 280 nm (left) and 295 nm (right). Top: data from the 1st analysis of the samples in ammonium bicarbonate buffer, at the received concentrations (~0.9

mg/mL). Bottom: data from the 2nd analysis, after 50- and 25-fold dilution of the biosimilar and the innovator samples respectively in PBS buffer, to a final concentration of 1 mg/mL.

The **2nd analysis** was aiming to assess whether the observed response at around 460 nm for the biosimilar samples 1–3 was still observed following dilution of the original formulations with PBS buffer. Raw fluorescence data (in three replicates) for the samples and the buffer are shown in Appendix B, Figure B18–Figure B19 for both excitation wavelengths (280 nm and 295 nm). The data following subtraction of the PBS signal are shown in Figure 4.16 (bottom). All six samples presented similarly shaped curves and none of them showed an additional peak at 460 nm. The maxima appeared at approximately 335.5 nm and 337.1 nm for excitation wavelengths 280 nm and 295 nm respectively (Table 4.11). The slightly shifted maxima (+0.6 nm) compared to the 1st analysis, especially with excitation at 280 nm, was complemented by the change in the calculated ratio of the intensities at 330 nm and 350 nm, which was decreased from 1.12 (1st analysis) to 1.11 (2nd analysis). This minor shift in the repeat analysis could perhaps have resulted from interactions of the protein with arginine, which was found to cause an expansion of the etanercept structure in the DLS results (Section 4.6.1) and could potentially have led to exposure of some of the Tyr/Trp residues to the solvent. The variation in the fluorescence intensities in PBS was within experimental error of the dilution.

Table 4.11: Key features of fluorescence emission spectra at excitation wavelengths 280 nm and 295 nm. The maxima wavelengths and the intensity ratio at 330:350 nm are given for the 1st and the 2nd analysis of the etanercept samples.

	Excitation at 280 nm				Excitation at 295 nm			
	Max (nm)		Intensity ratio 330:350 nm		Max (nm)		Intensity ratio 330:350 nm	
Sample	1st Analysis	2nd Analysis	1st Analysis	2nd Analysis	1st Analysis	2nd Analysis	1st Analysis	2nd Analysis
TNFR2:Fc 1	334.8	335.2	1.12	1.11	336.6	336.6	1.08	1.08
TNFR2:Fc 2	334.6	334.9	1.12	1.11	336.4	336.8	1.08	1.08
TNFR2:Fc 3	334.6	335.4	1.12	1.11	336.8	336.9	1.08	1.08
TNFR2:Fc 4	335.2	335.8	1.12	1.11	337.0	337.1	1.07	1.08
TNFR2:Fc 5	335.2	335.6	1.12	1.11	336.6	337.1	1.07	1.08
TNFR2:Fc 6	334.6	335.4	1.12	1.11	336.4	337.2	1.08	1.08

In conclusion, the observed response at 460 nm in the biosimilar samples from the 1st analysis was probably created during buffer exchange. The variation in the fluorescence intensities among samples in the 1st analysis was probably due to inaccuracy in the initially assigned concentrations, while the variation following dilution (for the 2nd analysis) was within the experimental error of the dilution. However, the key feature for assessing similarity in the environment of the aromatic residues is the emission wavelength at which the fluorescence intensity is maximal. This wavelength was consistent between samples within each analysis for both excitation wavelengths, while a small shift (+0.6 nm) was evident for the 2nd analysis with excitation at 280 nm, perhaps indicating changes in the environment of some Tyr/Trp residues due to interaction with buffer excipients (arginine). In all cases, the similarity of the emission wavelengths (334.6–337.2 nm) after exciting either at 280 nm or 295 nm, indicated dominance of tryptophan fluorescence even with excitation at 280 nm.

Taking into account the results from all four techniques applied, the etanercept samples appeared to be similar in the 1st analysis (far-/near-UV CD data at room temperature, FT-IR data and fluorescence collected for samples in ammonium bicarbonate buffer at the received concentration of ~0.9 mg/mL), with sample 1 showing some differences from the others for all techniques. However, the 2nd analysis—which included far-UV CD spectra before and after the melting curves, CD melting curves, all DLS data and repeat fluorescence, recorded for the samples diluted with PBS to 0.1 mg/mL (for far-UV CD) or 1 mg/mL (all other techniques)—showed all six samples to be the same within experimental error and equally stable upon heating. This led to the conclusion that the observed variation for sample 1 during the 1st analysis probably resulted from a concentration inaccuracy. DLS showed the presence of some larger particles in the biosimilar samples, but of insignificant population. During the temperature experiment, 55 °C seemed to be a key temperature, while at the end of the temperature experiment (cooled to 20 °C), etanercept appeared partially unfolded yet no aggregation was observed, perhaps due to the protective effect of ArgHCl against aggregation.[65] The CD melting curves showed similar paths of unfolding for all six samples and were in agreement with DLS showing that T_m1 was at 53 °C. The CD temperature experiment (recorded only for sample 1) showed that the fully unfolded conformation was reached at 80 °C. CD spectra recorded after cooling the samples down to 25 °C (following the melting curves) showed predominantly random coil structure and perhaps partial refolding of the protein, while

the respective HT spectra (before, during and following the melting curves) showed only insignificant signs of aggregation.

Overall, DLS showed similar sizes and stability for all six samples. The far- and near-UV CD spectra suggested that the samples had similar secondary and tertiary structure, within experimental and analysis error. The secondary structure analysis from the CD spectra of etanercept showed a large variation due to the unusual spectrum of the protein. However, the averaged values from all the performed analyses showed similar secondary structure contents between the biosimilar and the innovator samples, but probably not quantitatively representative of the actual structure. The CD melting curves suggested equal thermal stability of all samples. The FT-IR spectra were similar for all six samples, within experimental and data processing error. Analysis of the amide I IR band suggested similar secondary structure conformation, though the results from SSNN should be seen with caution (especially for α-helix and random coil). Finally, fluorescence showed similar environment for the Tyr/Trp residues of all six samples, within the experimental error of the dilution.

4.7. MONOCLONAL ANTIBODY IN PHASE 3

4.7.1. DLS spectroscopy of a monoclonal antibody

DLS data for the four monoclonal antibodies were collected by Dr. Claire E. Broughton. The samples were diluted to 5 mg/mL, using a buffer that was the same as the formulation buffer but omitting Tween 80. The hydrodynamic diameter of the samples was measured as 10.3–11.0 nm (Z-average value as calculated by the Zetasizer software, based on intensity values), with sample 1 presenting the smallest size and sample 4 the largest. Intensity values are usually closer to the absolute size of the particles, however, the presence of larger particles can affect the estimated size, as seems to be the case here. In fact, sample 4 presented a narrower averaged distribution (over 8 measurements), with higher intensity and a slightly smaller size than the other three samples, as can be observed from the intensity distributions (Figure 4.17, left). However, a larger variation was evident over the course of measurements for sample 4, suggesting a mixture of particles with different sizes (Figure B22, Appendix B). In all four samples, some large particles were present with a size of ~6000 nm (probably not related to the protein), which seemed to have a greater impact on the calculated Z-average value of sample 4, in which they appeared with a slightly higher intensity. All four samples showed moderate polydispersity with a PDI value of 0.2, therefore for the comparability purpose of this study the number values were used.

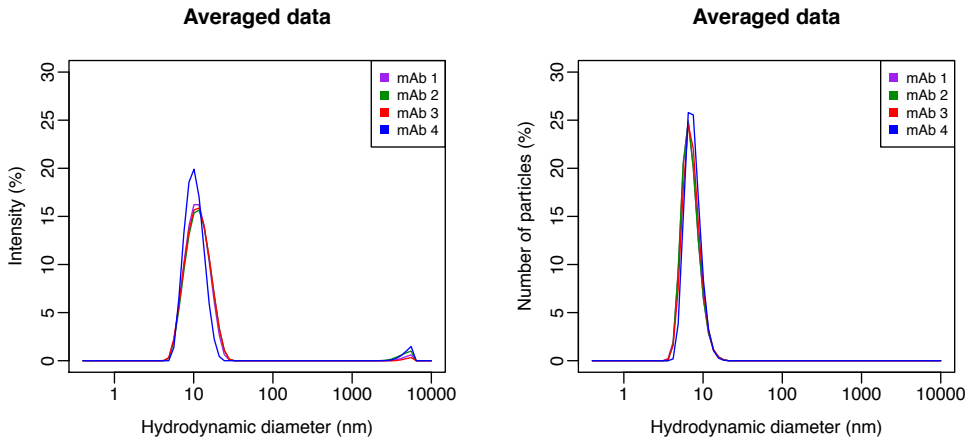
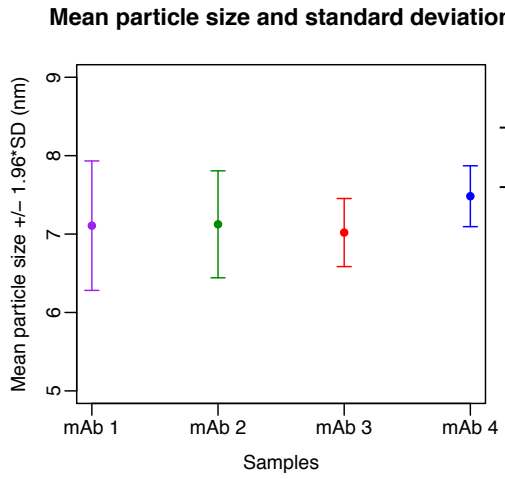


Figure 4.17: Left: Size distributions (nm) presented in intensity values for the four monoclonal antibody samples. Right: Size distributions (nm) of the four samples presented in number values. The presented data were recorded for the diluted samples to 5 mg/mL and are averaged over 8 measurements.

The distributions of the samples in number values (Figure 4.17, right) showed that the population of the large particles (~6000 nm) was not reproducible, samples 1–

3 presented identical distributions and sizes of 7.0–7.1 nm, while sample 4 presented a slightly larger average size of 7.5 nm and a peak with two maxima (at 6.5 and 7.5 nm). The mean sizes and standard deviations were calculated from the mean number values of each measurement (as calculated by the Zetasizer software) and are shown in Figure 4.18, with the exact values in the attached table.



Sample	Mean ± 1.96*SD
mAb 1	7.1±0.8
mAb 2	7.1±0.7
mAb 3	7.0±0.4
mAb 4	7.5±0.4

Figure 4.18: Left: Overall mean particle diameters (nm) of the four monoclonal antibody samples indicated as solid circles. For each sample, the overall mean was calculated over 8 measurements using the number mean values as estimated by the Zetasizer software. The error bars indicate the standard deviation with a 95% confidence interval. The values of the overall means and the standard deviations for each sample are shown in the attached table (right).

In conclusion, DLS showed similarly sized particles for samples 1–3, while sample 4 appeared somewhat different. The large variation throughout the measurements (intensity distributions), the two maxima and the slightly larger mean size (number values) of sample 4, probably suggest the presence of particles with different sizes.

4.7.2. CD spectra of a monoclonal antibody

CD spectra of the monoclonal antibody samples were recorded for the diluted samples at a concentration of ~0.1 mg/mL in 10 mM sodium phosphate buffer (pH ~7.0) using a 0.1 cm fixed path length cuvette. The four recorded baselines of the formulation buffer showed significant intensity signals at ~198 nm, probably due to arginine and citric acid excipients in the original formulation, even after 1000-fold dilution (Figure B23, Appendix B). The processed data (baseline-corrected and zeroed) are shown in Figure 4.19 (left). As already seen for trastuzumab in Chapter 3, all four samples presented the typical CD spectrum of an immunoglobulin with zero intensity at 208.2

nm and the typical shape of a β-sheet protein with a positive peak at 201 nm, a negative peak at 216 nm and a shoulder at 228 nm. Samples 1–3 appeared quite similar, indicating similarity in their secondary structure, while sample 4 appeared somewhat different in intensity but still of comparable secondary structure.

Spectra recorded at the end of the melting curves, after cooling down the samples to 25 °C (Figure 4.19, right), showed a broad negative peak at ~210 nm for all four samples, with sample 4 presenting a wider and slightly shifted peak, at ~212 nm. In all cases the spectra after returning to room temperature indicate a mixture of partially unfolded and completely unfolded molecules, as they seem to be a combination of a β-sheet and a random coil spectrum. Thermal denaturation, therefore, led to irreversibility due to the "off-pathway" events in the energy funnel, in the effort of the protein to obtain a conformation located in one of the neighbouring minima (see Section 1.2 and as described in Section 3.5.1). One of these minima was also represented by structures stabilised by hydrophobic clustering (aggregation) between the partially unfolded conformations. Looking back to Chapter 3 (Figure 3.30, bottom) the partially unfolded spectrum could be represented by the spectra of the native antibodies (mAbs) at 78 °C and the completely unfolded state illustrated by the spectra at 83 °C. The fact that sample 4 presents a significantly larger signal after cooling down is not completely related to its initially larger signal (at 25 °C), but is mostly due to dominance of the random coil spectrum (as this usually presents a stronger negative signal than β-sheet) and therefore the presence of a larger fraction of completely unfolded molecules compared to samples 1–3. This observation is in agreement with the DLS data which indicated the presence of differently sized particles in sample 4, perhaps suggesting that a fraction of the protein molecules was in a partially unfolded conformation from the beginning.

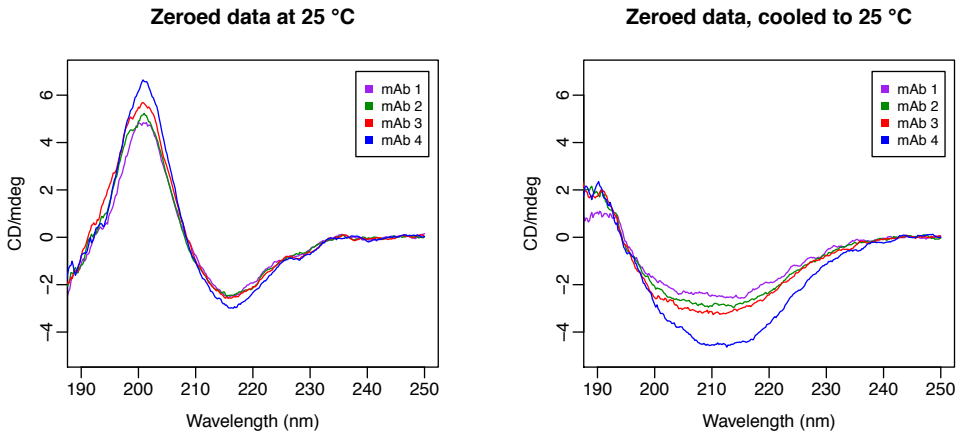


Figure 4.19: Comparison of the CD spectra (mdeg) of the diluted monoclonal antibody samples at a concentration of ~0.1 mg/mL using a 0.1 cm fixed path length cuvette. Spectra were acquired before recording the melting curves (25 °C, left), and at the end of the temperature experiment after the samples were cooled down to 25 °C (right).

Melting curves

The signals of the maximum at 200.0 nm and of the minimum at 217.0 nm were followed for all four samples throughout a range of temperatures from 25 °C to 100 °C resulting in the melting curves shown in Figure 4.20 (left and right respectively). All four samples presented a common path of unfolding at 200 nm, obtaining their completely unfolded conformation at ~78 °C, while sample 4 showed the largest change in the intensity of its signal and sample 1 the smallest, which were also related to their respective intensities for their 25 °C spectra. Decrease of the intensity at 200 nm indicates either an increase in the random coil content or a decrease in α-helix content, as the wavelength 200 nm is affected by the characteristic spectra of both conformations. The curves at 217 nm were very similar for samples 1–3 presenting the onset of loss of β-sheet structure at around 58 °C, while at ~73 °C β-sheet appeared to increase and was stabilised at ~78 °C. Increase in the β-sheet content perhaps suggests that aggregation took place, as aggregation usually occurs for partially unfolded proteins after interaction of their β-sheet segments.[73,74] Sample 4 appeared significantly different than the rest, showing the lowest β-sheet content, at which it was stabilised at ~75 °C. Visual inspection of all samples at the end of the melting curves measurements and after cooling down to room temperature, confirmed the presence of solid particles in all samples confirming aggregation, perhaps favoured at the used pH, as neutral pH conditions have been found in literature to increase aggregation of antibodies.[75,76] The raw melting curves for both wavelengths are shown in Figure B24 (Appendix B).

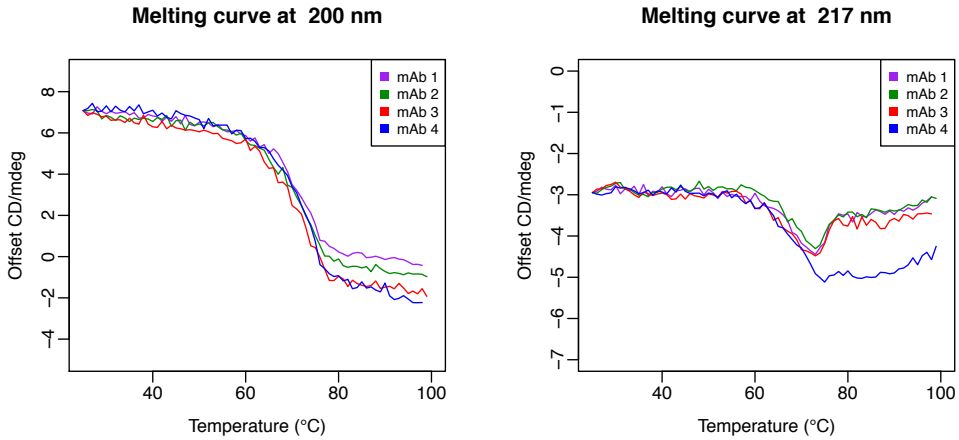


Figure 4.20: Offset (as described in materials and methods, Section 4.2.4) melting curves (mdeg) of the monoclonal antibody samples at 200.0 nm (left) and 217.0 nm (right). The curves were recorded for the samples diluted with sodium phosphate buffer, resulting in a final solution of ~0.1 mg/mL protein in 10 mM sodium phosphate buffer. The 0.1 cm fixed path length cuvette was used.

In conclusion, samples 1–3 presented quite similar spectra, and therefore secondary structure, while sample 4 appeared to be of comparable secondary structure based on its CD spectrum compared to the spectra of samples 1–3. All four samples were stable up to 58 °C. Samples 1–3 were of equal stability perhaps showing signs of aggregation of a partially unfolded conformation at increased temperature, while sample 4 presented lower β-sheet content and a completely unfolded conformation at 75 °C. After cooling down, the samples presented a mixture of partially unfolded and completely unfolded molecules, with sample 4 appearing slightly different, perhaps due to the presence of a larger fraction of completely unfolded molecules compared to samples 1–3. Solid particles, and consequently aggregation, were obvious in all samples following visual inspection at the end of the measurements.

4.7.3. 1D ^{1}H-NMR spectroscopy of a monoclonal antibody

1D ^{1}H-NMR spectra for the four monoclonal antibodies were recorded at 45 °C, after the samples had been buffer exchanged into sodium phosphate buffer at pH 5.4—as the protein is more stable at pH 4.5–5.5 and the original formulation was at pH 5.0—and at concentrations of approximately 30–45 mg/mL. The used temperature (45 °C) has been previously mentioned in the literature to enhance homogeneity of the samples,[77,78] while the secondary structure of the protein is not expected to be affected since antibodies are usually stable at least up to 50 °C, which was also indicated by the CD melting curves. In order to improve the quality of the spectra, the experimental time was significantly longer compared to the one required for insulin (2 h compared to 5

min), as the molar concentration of the antibody was smaller (~200 μM compared to 1 mM for insulin). Exhaustive buffer exchange was necessary as the signals of the formulation excipients were very strong in the original samples, covering the protein signals. Care was required in order to ensure the same conditions for all samples, as small variations in pH can result in major spectral differences and mislead the conclusions. Absence of stabilising excipients, and especially the surfactant Tween 80, in such high concentrations might favour formation of aggregates. However, this did not seem to be the case here, as further widening of the peaks would have been expected given the size of the protein. This is also supported by the 1D ^{1}H-NMR spectra recorded in sodium phosphate buffer at neutral pH (7.0), which presented broadening in their line widths compared to the spectra of the samples at pH 5.4 (Figure B25, Appendix B). The fact that the 1D ^{1}H-NMR spectra at neutral pH were recorded before the spectra at pH 5.4 perhaps suggests that reversible aggregation was favoured at neutral pH and absence of stabilising excipients.

The 1D NMR spectra for the four samples and the formulation buffer on its own are shown in Figure 4.21, where it is obvious that only insignificant traces of the formulation buffer were present in the samples. However, a few sharp peaks appeared, that were probably impurities induced during sample preparation, and were removed from the spectra for clarity, as well as the water peak. An overall look at the four antibody spectra does not disclose any distinct differences among the samples, except for a slightly weaker intensity of sample 4 compared to the rest.

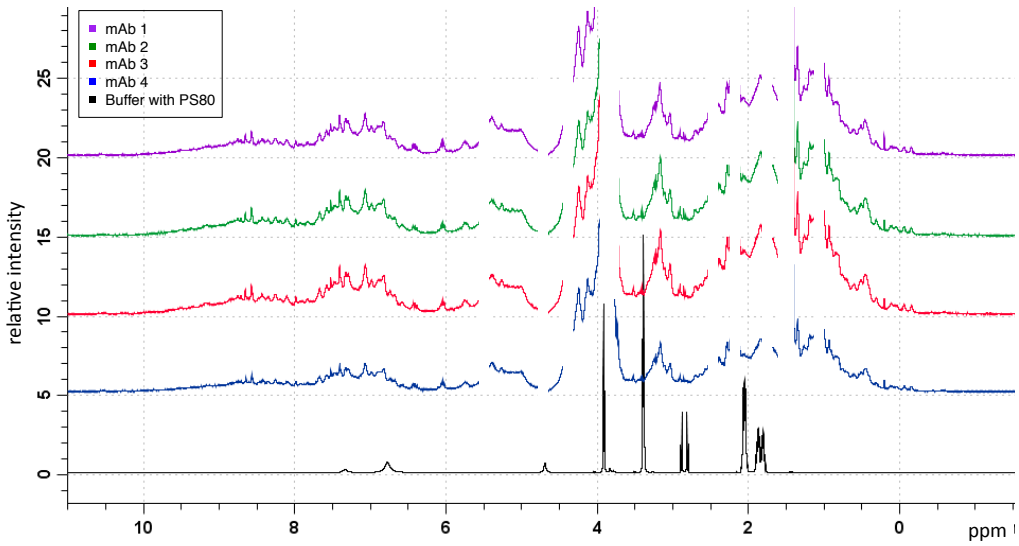


Figure 4.21: 1D ^{1}H-NMR spectra of the four monoclonal antibody samples, after buffer exchange into 20 mM sodium phosphate buffer, at a protein concentration of approximately 30–45 mg/mL and pH 5.4. Samples 1–4 are shown in purple, green, red and blue respectively. In black is the formulation buffer spectrum. All spectra were recorded at 45 °C.

Subtracting sample 2 from sample 1 (Figure 4.22) indicated a few differences in the regions 6.0–8.0 ppm (sidechain H^N, aromatic protons and backbone H^N), 4.0–4.3 ppm (Hα resonances), 2.7–3.5 ppm and 0.3–1.3 ppm (aliphatic protons). Samples 2 and 3 appeared almost identical (Figure 4.23), with only minor differences in the Hα and aliphatic region, at 4.0–4.3 ppm and 2.7–3.5 ppm respectively. Samples 2 and 4 (Figure 4.24) presented some variation in the aromatic/H^N area (6.0–9.0 ppm), in the Hα region (4.0–4.3 ppm) and in the aliphatic region (2.5–3.5 ppm and 1.2–1.5 ppm). Finally, samples 1 and 4 (Figure 4.25) varied in the aromatic/H^N (6.3–9.0 ppm), Hα (4.0–4.3 ppm) and aliphatic region (2.7–3.5 ppm and 0.3–1.3 ppm). Variation in the regions 4.0–4.3 ppm and 2.7–3.5 ppm, that was present in all subtractions, might result from heterogeneity in glycosylation, as usually glycan signals appear at around 3.0–5.0 ppm. Therefore, it cannot be inferred with certainty whether differences in these areas are due to lack of similarity in the higher-order structure of the samples or associated with the present glycan patterns.

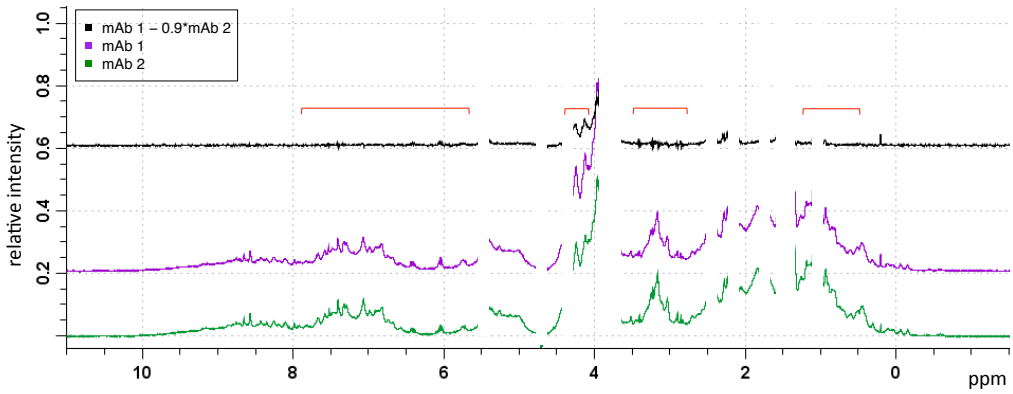


Figure 4.22: Residual ^{1}H-NMR spectrum (black) after subtracting the spectrum of sample 2 (green) from sample 1 (purple). A scaling factor of 0.9 was used for the subtraction. Red brackets indicate the regions of the spectra that appeared to vary after subtraction.

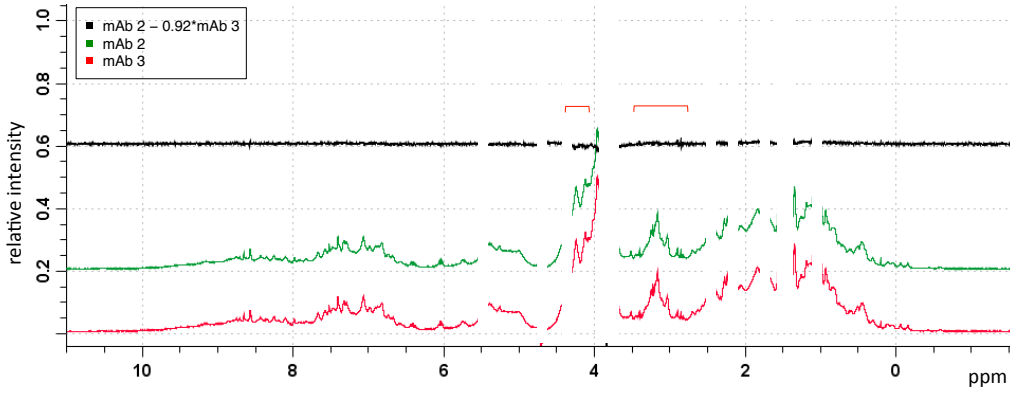


Figure 4.23: Residual ^{1}H-NMR spectrum (black) following subtraction of sample 3 (red) from sample 2 (green). A scaling factor of 0.92 was used for the subtraction. Red brackets indicate the regions of the spectra that appeared to vary after subtraction.

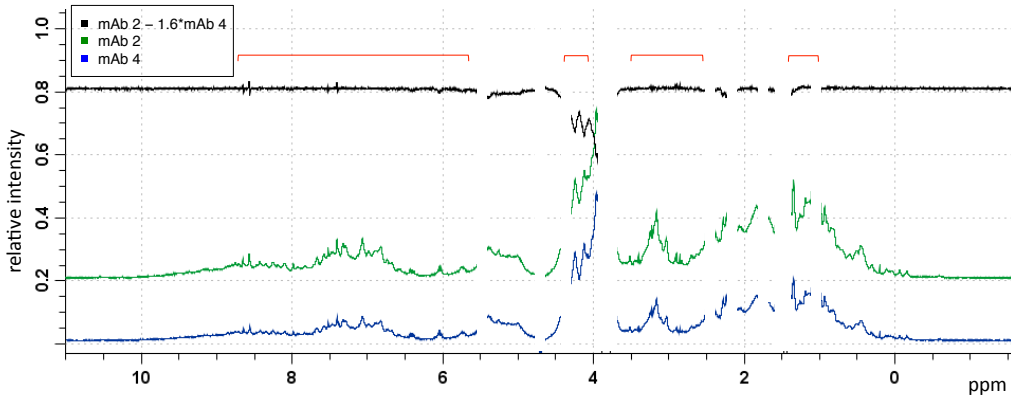


Figure 4.24: Residual ^{1}H-NMR spectrum (black) following subtraction of sample 4 (blue) from sample 2 (green). A scaling factor of 1.6 was used. Red brackets indicate the regions of the spectra that appeared to vary after subtraction.

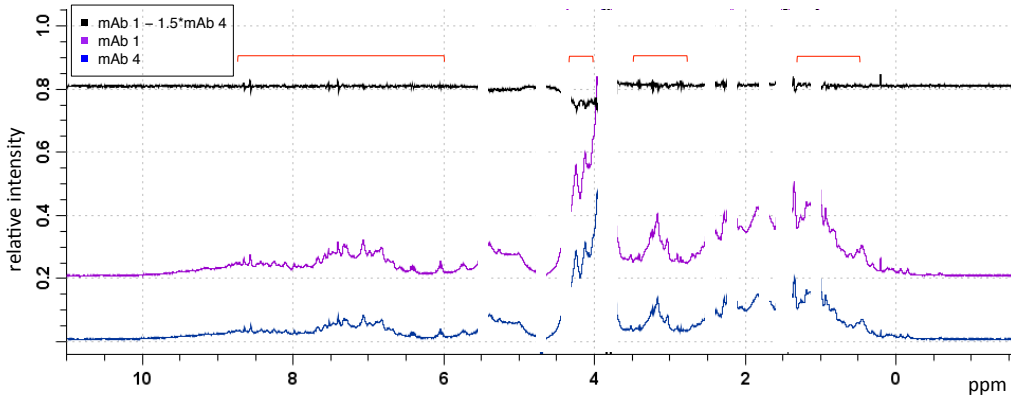


Figure 4.25: Residual ^{1}H-NMR spectrum (black) following subtraction of sample 4 (blue) from sample 1 (purple). A scaling factor of 1.5 was used. Red brackets indicate the regions of the spectra that appeared to vary after subtraction.

In conclusion, the 1D ^{1}H-NMR spectra of the four monoclonal antibody samples showed similarity between samples 2 and 3, while samples 1 and 4 appeared somewhat different from each other and from samples 2 and 3. Overall, all three techniques (CD, DLS and NMR) agreed on similarity of samples 2 and 3 and variation of sample 4, while differences in sample 1 were only evident in the NMR spectra, as its particles size, secondary structure and thermal stability were found to be almost identical to samples 2 and 3. However, it should be kept in mind that the different concentrations used in each technique and the varied pH ranges (7.0 or 5.0–5.4) and buffer conditions (presence or absence of excipients) might have induced various structural changes, to which NMR is very sensitive.

4.8. CONCLUSIONS

In the cases of protein H and hGH, only far-UV CD was conducted for comparability studies and investigation of their secondary structure respectively. Though CD is sufficient and the most direct technique for probing the secondary structure of proteins, it cannot show with certainty variation in particle populations and is not sensitive to differences at amino acid level. For these purposes, complementarity with DLS and near-UV CD or fluorescence spectroscopy is necessary. Protein A has an unusual far-UV CD spectrum of β_{II} type, which is hard to differentiate from a random coil spectrum. Near-UV CD was therefore conducted in addition, in order to prove the presence of ordered structure in the protein. Etanercept has also an unusual far-UV CD spectrum, with predominant random coil and β-sheet conformation. The study in this case was complemented by FT-IR for additional confirmation on the secondary structure of the protein; near-UV CD and fluorescence looked into the tertiary structure and orientation of the aromatic amino acids, while DLS revealed that the innovator samples were of higher purity due to the presence of some larger particles in the biosimilar samples. Moreover, coupling of DLS with temperature-dependent measurements and the CD melting curves showed similar stability for all etanercept samples under stressed conditions. For the monoclonal antibody study (mAb), DLS already indicated particle size variation in one of the samples (sample 4), which perhaps could also explain for the different far-UV CD spectrum compared to samples 1–3. CD melting curves showed differences in stability of sample 4 compared to the remaining three samples. In addition, 1D ^{1}H-NMR confirmed the observed variation of sample 4, while it also revealed some differences for sample 1, which were not obvious from CD or DLS. However, since NMR is quite sensitive to differences in atom level, all four samples showed some variation in the area 3.0–5.0 ppm, which might result from the glycosylation patterns and not from actual changes in the higher-order structure of the protein.

As far as the encountered challenges are concerned, secondary structure analysis of the far-UV CD spectra using different algorithms and reference sets appeared to vary significantly in some cases, because of two reasons: unusual spectrum of the examined protein and intensity error accumulated from different sources. Combination of analyses with various software and reference sets usually offers an improved representation of the under-analysis protein, while comparison of the results with literature or the DSSP assignment for the PDB crystal structure is advisable when more accurate structure percentages are required. The studies in the present chapter

showed that an intensity error may occur mainly due to high viscosity of concentrated samples, which can result in a path length increase up to 50% when a demountable cuvette is used. Concentration errors (*e.g.* due to solvent evaporation or human error at sample dilution) and intensity calibration deviation can contribute up to ~5% and ~8% respectively, and can either cancel out or add up to a path length error. However, the often significant path length error was addressed to a satisfactory level by using a fixed cuvette, the instrument calibration error can be eliminated by applying a scaling factor, while concentration errors are harder to control. Finally, additional sample preparation such as dilution of highly concentrated samples (*e.g.* for CD) or buffer exchange due to the unsuitable formulation buffer, depending on the used analytical method (*e.g.* CD or NMR) is occasionally necessary (*e.g.* mAb study), though not ideal as it might induce changes in the secondary structure of the protein or impurities. If dilution is not with the formulation buffer errors in buffer component concentrations can cause further errors. Such was potentially the case for protein A, where the secondary structure of the protein changed upon dilution and the mAb study, where buffer exchange caused induction of impurities in the samples. However, as long as all samples were treated similarly, sample-to-sample comparison was still possible within each technique and combination of the results confirmed similarity or variation for protein A and mAb respectively. Complementarity of the used techniques in order to draw conclusions on similarity among samples is discussed in Chapter 5.

References

1. W. S. Evans, A. C. Faria, E. Christiansen, K. Y. Ho, J. Weiss, A. D. Rogol, M. L. Johnson, R. M. Blizzard, J. D. Veldhuis and M. O. Thorner, *American Journal of Physiology - Endocrinology And Metabolism*, 1987, **252**, E549-E556.
2. J. C. Beck, E. E. McGarry, I. Dyrenfurth and E. H. Venning, *Science (Washington)*, 1957, **125**, 884-885.
3. M. S. Raben, *Science*, 1957, **125**, 883-884.
4. N. Mauras, *Trends in Endocrinology & Metabolism*, **8**, 256-257.
5. S. J. Kim, S. K. Hahn, M. J. Kim, D. H. Kim and Y. P. Lee, *Journal of Controlled Release*, 2005, **104**, 323-335.
6. J. L. Cleland, N. C. Geething, J. A. Moore, B. C. Rogers, B. J. Spink, C.-W. Wang, S. E. Alters, W. P. C. Stemmer and V. Schellenberger, *Journal of Pharmaceutical Sciences*, 2012, **101**, 2744-2754.
7. F. Fares, R. Guy, A. Bar-Ilan, Y. Felikman and E. Fima, *Endocrinology*, 2010, **151**, 4410-4417.

8. K. Sprogoe, M. Beckert, E. D. Christoffersen, D. Gilfoyle and T. Wegge, in *GH-Pediatrics (posters)*, Endocrine Society, 2016, pp. SAT-035-SAT-035.
9. B. L. Osborn, L. Sekut, M. Corcoran, C. Poortman, B. Sturm, G. Chen, D. Mather, H. L. Lin and T. J. Parry, *European Journal of Pharmacology*, 2002, **456**, 149-158.
10. M. H. Rasmussen, M. W. B. Olsen, L. Alifrangis, S. Klim and M. Suntum, *The Journal of Clinical Endocrinology & Metabolism*, 2014, **99**, E1819-E1829.
11. E. S. Kim, D. S. Jang, S. Y. Yang, M. N. Lee, K. S. Jin, H. J. Cha, J. K. Kim, Y. C. Sung and K. Y. Choi, *Nanoscale*, 2013, **5**, 4262-4269.
12. A. Colao, G. Vitale, R. Pivonello, A. Ciccarelli, C. Di Somma and G. Lombardi, *European Journal of Endocrinology*, 2004, **151**, S93-101.
13. G. Lombardi, A. Colao, P. Marzullo, D. Ferone, S. Longobardi, V. Esposito and B. Merola, *Journal of Endocrinology*, 1997, **155**, S33-S37.
14. O. L. Johnson, W. Jaworowicz, J. L. Cleland, L. Bailey, M. Charnis, E. Duenas, C. Wu, D. Shepard, S. Magil, T. Last, A. J. S. Jones and S. D. Putney, *Pharmaceutical Research*, 1997, **14**, 730-735.
15. Firstword Pharma, Genentech to discontinue Nutropin Depot, https://www.firstwordpharma.com/node/217255, (accessed 19 Apr, 2017).
16. S. Aloj and H. Edelhoch, *Journal of Biological Chemistry*, 1972, **247**, 1146-1152.
17. L. A. Holladay, R. G. Hammonds and D. Puett, *Biochemistry*, 1974, **13**, 1653-1661.
18. A. M. de Vos, M. Ultsch and A. A. Kossiakoff, *Science*, 1992, **255**, 306.
19. T. A. Bewley and C. H. Li, *Biochemistry*, 1972, **11**, 884-888.
20. R. Pearlman and T. A. Bewley, in *Stability and Characterization of Protein and Peptide Drugs: Case Histories*, eds. Y. J. Wang and R. Pearlman, Springer US, Boston, MA, 1993, DOI: 10.1007/978-1-4899-1236-7_1, pp. 1-58.
21. L. Chantalat, N. Jones, F. Korber, J. Navaza and A. Pavlovsky, *Protein and Peptide Letters*, 1995, **2**, 333-340.
22. M. M. Goldenberg, *Clinical Therapeutics*, 1999, **21**, 75-87.
23. Marc Feldmann, a. Fionula M. Brennan and Ravinder N. Maini, *Annual Review of Immunology*, 1996, **14**, 397-440.
24. FDA, Product Approval Information - Licensing Action, https://www.accessdata.fda.gov/drugsatfda_docs/appletter/1998/etanimm110298L.htm, (accessed 5 May, 2017).
25. EMA, Enbrel - Authorisation details, http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/medicines/000262/human_med_000764.jsp&mid=WC0b01ac058001d124, (accessed 6 May, 2017).
26. N. Watanabe, H. Kuriyama, H. Sone, H. Neda, N. Yamauchi, M. Maeda and Y. Niitsu, *Journal of Biological Chemistry*, 1988, **263**, 10262-10266.
27. H. Lee, H. C. Kimko, M. Rogge, D. Wang, I. Nestorov and C. C. Peck, *Clinical Pharmacology & Therapeutics*, 2003, **73**, 348-365.
28. E. D. Lobo, R. J. Hansen and J. P. Balthasar, *Journal of Pharmaceutical Sciences*, 2004, **93**, 2645-2668.

29. V. F. Azevedo, N. Galli, A. Kleinfelder, J. D'Ippolito and P. C. M. Urbano, *Rheumatology International*, 2015, **35**, 197-209.
30. M. DiPaola, J. Li and E. Stephens, *Journal of Bioanalysis & Biomedicine*, 2013, **5**, 180-186.
31. S. M. Mahler, A. Wardiana, M. L. Jones, C. J. de Bakker, G. G. Graham and C. B. Howard, *International Journal of Rheumatic Diseases*, 2016, **19**, 1043-1048.
32. EMA, Benepali - Authorisation details, http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/medicines/004007/human_med_001944.jsp&mid=WC0b01ac058001d124, (accessed 6 May, 2017).
33. FDA, FDA approves Erelzi, a biosimilar to Enbrel, https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm518639.htm, (accessed 3 May, 2017).
34. EMA, Lifmior - Authorisation details, http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/medicines/004167/human_med_002080.jsp&mid=WC0b01ac058001d124, (accessed 6 May, 2017).
35. S. Houel, M. Hilliard, Y. Q. Yu, N. McLoughlin, S. M. Martin, P. M. Rudd, J. P. Williams and W. Chen, *Analytical Chemistry*, 2014, **86**, 576-584.
36. C. S. Sevier and C. A. Kaiser, *Nature Reviews Molecular Cell Biology*, 2002, **3**, 836-847.
37. I. H. Cho, N. Lee, D. Song, S. Y. Jung, G. Bou-Assaf, Z. Sosic, W. Zhang and Y. Lyubarskaya, *mAbs*, 2016, **8**, 1136-1155.
38. Y. Mukai, T. Nakamura, M. Yoshikawa, Y. Yoshioka, S. Tsunoda, S. Nakagawa, Y. Yamagata and Y. Tsutsumi, *Science signaling*, 2010, **3**, ra83.
39. S. Matsumiya, Y. Yamaguchi, J. I. Saito, M. Nagano, H. Sasakawa, S. Otaki, M. Satoh, K. Shitara and K. Kato, *Journal of Molecular Biology*, 2007, **368**, 767-779.
40. A. Ceroni, A. Dell and S. M. Haslam, *Source Code for Biology and Medicine*, 2007, **2**, 3.
41. D. Damerell, A. Ceroni, K. Maass, R. Ranzinger, A. Dell and S. M. Haslam, *Biological Chemistry*, 2012, **393**, 1357-1362.
42. S. Krishnan, M. M. Pallitto and M. S. Ricci, in *Formulation and Process Development Strategies for Manufacturing Biopharmaceuticals*, John Wiley & Sons, Inc., 2010, DOI: 10.1002/9780470595886.ch16, pp. 383-427.
43. T. Arvinte and P. Fauquex, Stable liquid formulation of antibodies, https://www.google.com/patents/US20040170623, (accessed 6 May, 2017).
44. J. Liu, M. D. H. Nguyen, J. D. Andya and S. J. Shire, *Journal of Pharmaceutical Sciences*, **94**, 1928-1940.
45. L. O. Narhi, C. H. Li, R. Ramachander, J. Svitel and Y. Jiang, in *Molecular Biophysics for the Life Sciences*, eds. N. Allewell, L. O. Narhi and I. Rayment, Springer New York, New York, NY, 2013, DOI: 10.1007/978-1-4614-8548-3_3, pp. 33-90.
46. J. R. Lakowicz, *Principles of Fluorescence Spectroscopy*, Kluwer Academic/Plenum, 1999.

47. T. L. Hwang and A. J. Shaka, *Journal of Magnetic Resonance, Series A*, 1995, **112**, 275-279.
48. J. P. Hennessey and W. C. Johnson, *Analytical Biochemistry*, 1982, **125**, 177-188.
49. S. Ludvigsen, M. Roy, H. Thoegersen and N. C. Kaarsholm, *Biochemistry*, 1994, **33**, 7998-8006.
50. N. Sreerama and R. W. Woody, *Analytical Biochemistry*, 2000, **287**, 252-260.
51. W. Kabsch and C. Sander, *Biopolymers*, 1983, **22**, 2577-2637.
52. W. C. Johnson, *Proteins: Structure, Function, and Bioinformatics*, 1999, **35**, 307-312.
53. S. M. King and W. C. Johnson, *Proteins: Structure, Function, and Bioinformatics*, 1999, **35**, 313-320.
54. N. Sreerama and R. W. Woody, *Biochemistry*, 1994, **33**, 10022-10025.
55. T. A. Bewley, J. Brovetto-Cruz and C. H. Li, *Biochemistry*, 1969, **8**, 4701-4708.
56. S. M. Kelly, T. J. Jess and N. C. Price, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 2005, **1751**, 119-139.
57. Y.-H. Chen, J. T. Yang and K. H. Chau, *Biochemistry*, 1974, **13**, 3350-3359.
58. L. Zhong and W. C. Johnson, *Proceedings of the National Academy of Sciences*, 1992, **89**, 4462-4465.
59. A. H. Fradkin, J. F. Carpenter and T. W. Randolph, *Journal of Pharmaceutical Sciences*, 2009, **98**, 3247-3264.
60. N. A. Kim, D. G. Lim, J. Y. Lim, K. H. Kim and S. H. Jeong, *International Journal of Pharmaceutics*, 2014, **460**, 108-118.
61. A. Hawe, W. L. Hulse, W. Jiskoot and R. T. Forbes, *Pharmaceutical Research*, 2011, **28**, 2302-2310.
62. T. Kohno, L. T. T. Tam, S. R. Stevens and J. S. Louie, *Journal of Investigative Dermatology Symposium Proceedings*, 2007, **12**, 5-8.
63. Malvern, Size: What is the Z-average?, http://www.malvern.com/en/support/resource-center/faq/FAQ0015AverageDiameter.aspx, (accessed 3 May, 2017).
64. T. Arakawa, D. Ejima, K. Tsumoto, N. Obeyama, Y. Tanaka, Y. Kita and S. N. Timasheff, *Biophysical Chemistry*, 2007, **127**, 1-8.
65. N. A. Kim, S. Hada, R. Thapa and S. H. Jeong, *International Journal of Pharmaceutics*, 2016, **513**, 26-37.
66. V. Medina Martinez, M. E. Abad-Javier, A. J. Romero-Diaz, F. Villasenor-Ortega, N. O. Perez, L. F. Flores-Ortiz and E. Medina-Rivero, *Journal of Analytical Methods in Chemistry*, 2014, **2014**, 11.
67. Z. Shi, R. W. Woody and N. R. Kallenbach, *Advances in Protein Chemistry*, 2002, **62**, 163-240.
68. A. Micsonai, F. Wien, L. Kernya, Y. H. Lee, Y. Goto, M. Réfrégiers and J. Kardos, *Proceedings of the National Academy of Sciences*, 2015, **112**, E3095-E3103.
69. D. M. Byler and H. Susi, *Biopolymers*, 1986, **25**, 469-487.
70. E. Goormaghtigh, V. Cabiaux and J.-M. Ruysschaert, in *Physicochemical Methods in the Study of Biomembranes*, eds. H. J. Hilderson and G. B. Ralston,

Springer US, Boston, MA, 1994, DOI: 10.1007/978-1-4615-1863-1_10, pp. 405-450.
71. A. Barth, *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 2007, **1767**, 1073-1101.
72. M. R. Eftink, in *Topics in Fluorescence Spectroscopy: Volume 6: Protein Fluorescence*, ed. J. R. Lakowicz, Springer US, Boston, MA, 2000, DOI: 10.1007/0-306-47102-7_1, pp. 1-15.
73. J. F. Carpenter, B. S. Kendrick, B. S. Chang, M. C. Manning and T. W. Randolph, *Methods in Enzymology*, 1999, **309**, 236-255.
74. E. Y. Chi, S. Krishnan, T. W. Randolph and J. F. Carpenter, *Pharmaceutical Research*, 2003, **20**, 1325-1336.
75. J. M. R. Moore, T. W. Patapoff and M. E. M. Cromwell, *Biochemistry*, 1999, **38**, 13960-13967.
76. Á. Szenczi, J. Kardos, G. A. Medgyesi and P. Závodszky, *Biologicals*, 2006, **34**, 5-14.
77. L. Poppe, J. B. Jordan, K. Lawson, M. Jerums, I. Apostol and P. D. Schnier, *Analytical Chemistry*, 2013, **85**, 9623-9629.
78. L. Poppe, J. B. Jordan, G. Rogers and P. D. Schnier, *Analytical Chemistry*, 2015, **87**, 5539-5545.

Chapter 5
Conclusion

Seven studies were completed in total in the present work, each one to a different extent, but with the same purpose: structural and physical characterisation of protein biotherapeutics. Such complex molecules require a series of tests in order to support comparability between a follow-on version (biosimilar) and the original (innovator) product, as well as batch-to-batch similarity. Changes in the used conditions in any step of the manufacturing process can induce heterogeneity, which can affect the stability, potency and immunogenicity of the final product. Full control of the manufacturing process is therefore of utmost importance, and accordingly significant is the inspection of the structural similarity and the amount of impurities or degradation products among samples. Far-UV CD spectroscopy was the common technique in all studies, while a number of analytical techniques such as: near-UV CD, DLS, FT-IR, fluorescence, 1D ^{1}H-NMR and MS were used in different combinations in each study, in order to identify what information can be extracted and how they complement each other. Temperature-dependent measurements were applied in some of the studies, for further investigations on the thermal stability of the protein.

What information does each technique offer and which are the challenges?

Mass spectrometry (MS) includes a number of sophisticated experiments for protein characterisation and is probably the most robust technique and currently the method of choice in comparability studies for biopharmaceutical products. In this work, MS was used as a tool to confirm the primary structure (amino acid sequence) of the antibody samples (mAb-ADC study, Chapter 3) and also in order to look at post-translational modifications, such as carbohydrate structure and position, and degradation products (asparagine deamidation and methionine oxidation), while relative quantitation of the toxin molecules on each one of the conjugated samples was also applied. However, MS is a very expensive technique both because of the cost of the instrument itself, as this also needs to be coupled with an HPLC separation system, and due to the pricey reagents (*e.g.* enzymes) that are usually required for treatment of the samples prior to analysis. Furthermore, MS experiments produce a very large amount of data, which is time-consuming to analyse even with the use of the available automated libraries.

The rest of the techniques that were used in these studies are in their majority non-invasive. They can usually be applied to the provided sample in its original solution, without requiring any additional sample preparation, with some exceptions for high-concentrated samples (50–100 mg/mL) or in cases where the formulation buffer is unsuitable for the chosen method. CD, DLS, FT-IR, fluorescence and 1D ^{1}H-NMR spectroscopies all offer the convenience of speed as far as acquiring, processing and interpretation of the results are concerned, while all of them apart from NMR have the additional advantage of being significantly cheaper techniques compared to MS.

Far-UV CD is the most direct technique for providing information on the secondary structure of a protein. However, often-encountered challenges might include intensity errors resulting from high viscosity of concentrated samples when used with a demountable cuvette, especially in the absence of stabilising excipients such as surfactants and salts which change the ionic strength of the solution and therefore decrease viscosity. In such cases, dilution of the samples might be necessary, however it should be made sure that the secondary structure of the protein is not affected by changes in concentration. An alternative solution to variations in the path length of a demountable cuvette due to high viscosity of the samples would be either the manufacturing of a sealed holder that will allow assembly of the cell ensuring a stable path length and also heating of the cell without the possibility of solvent evaporation or leakage, or the use of a sealed circular cell of the same path length as the demountable cuvette—both already in use by Prof. Bonnie Wallace. Visual inspection of the CD spectra is usually the first step to assess any observed differences or errors, while quantitation of the results can be achieved by the use of secondary structure analysis software. Combination of various programs and reference sets often offers better comparison, especially when comparing results from the same software/reference set for each sample, as in many cases different algorithms and reference sets present significant variation. This variation usually results from differences in the performance of each algorithm (*e.g.* CDSSTR consistently shows smaller spectral error than SELCON3, without this meaning that the quality of the analysis is always better) or from an unusual CD spectrum or even from accumulation of an intensity error. SSNN provided good comparative results for the temperature-dependent experiments for insulin (Chapter 2) and mAb-ADC (Chapter 3) studies.

FT-IR is also a useful technique for investigating the secondary structure of proteins, as an alternative to CD, while it is probably the most effective spectroscopic method when it comes to fibrillation studies. However, processing of the FT-IR spectra

and more specifically buffer and water vapour corrections can result in variability in the shape of the amide I band when dealing with small concentrations (~1 mg/mL). Higher concentrations (≥10 mg/mL) are preferable when this is feasible, while the great flexibility of the technique allows for the analysis of solid samples and deposited layers of liquid samples after they dry out (as applied in the case of insulin). Nevertheless, the latter might not be always desirable as in most cases the structure of the protein in solution is of main interest.

DLS is the most straightforward technique to provide information on the purity and homogeneity of the analysed sample. Particle size is a fundamental physical characteristic of a globular protein, while variation in size can indicate the presence of different populations and therefore changes in conformation, oligomerisation and aggregation state. It is one of the most versatile techniques as it can deal with a wide variety of concentrations as well as formulation buffers.

Near-UV CD provides additional information on protein folding, looking at the tertiary structure of the molecule and more specifically at conformational changes in the environment of the aromatic amino acid side chains and disulfide bonds. Each protein has a unique near-UV CD spectrum, characteristic for the specific molecule as this can differ even among the same protein from different species. Weakening of the CD signal in the near-UV region usually indicates loss of ordered structure. Its sensitivity makes it a very useful tool in comparability studies and a reliable complement to far-UV CD data.

Fluorescence emission spectra are also sensitive to changes in tertiary structure and orientation of the aromatic amino acids, especially tryptophan and tyrosine. Changes in intensity occur due to conformational changes that alter interactions between the aromatic amino acid side chains and nearby quenchers, while shift of the maximum towards longer wavelengths indicates unfolding of the structure.

NMR is useful when dealing with high concentrated samples. It offers more detailed information compared to the above-mentioned techniques, and is able to point out subtle changes in the higher-order structure of the protein that might be missed by the rest of the methods. However, the exact same conditions should be ensured for all samples, as slight changes in pH can result in different chemical shifts. Furthermore, the formulation buffer is an additional limiting factor as the commonly used excipients overlap with the protein signals. In such cases buffer exchange might be necessary, though not desirable, as it can cause conformational and stability changes to the protein as well as induce impurities. NMR spectra are somewhat hard to interpret due to the

difficulty in specifying whether the observed variation results from intrinsic protein heterogeneity (such as glycan patterns) or from differences in the higher-order structure of the protein.

Solving the puzzle: Complementarity of Techniques in drawing conclusions for the probed proteins

Application of different spectroscopic techniques as complementary to each other provides undoubtedly a more complete idea on the similarity of the assessed samples, since each of the selected methods investigates different biophysical phenomena. However, some of the techniques provide answers to similar questions. Far-UV CD and FT-IR, for example, both examine the secondary structure of the molecule, and similarly near-UV CD and fluorescence both provide information on tertiary structure and orientation of the aromatic amino acid side chains. Combination of the correct techniques is, therefore, not a matter of which technique can be de facto omitted or replaced by another, but what information is required in each case. The flexibility of each method, as far as the used concentration or formulation buffer are concerned, can be used in favour of the scientist, provided that the experiments are planned accordingly and depending on the under-analysis protein and its formulation.

In any study, the academic way imposes the importance that the researcher should be informed on literature, when published data are available. However, industry works in much shorter time-scales and is therefore interested in obtaining results as fast as possible, ideally performing a series of **"blind experiments"** that would however ensure that the desired conclusions will be reached. The conducted studies in the present work have provided empirical experience, which can offer practical advice when decisions should be made on which of the above biophysical techniques should be used and in which cases. A potential "SOP" would suggest the following:

> Given the fact that MS is the only technique to provide detailed information on the **primary structure**, **post-translational modifications** and degradation products, it is still the basis of the quality control tests.
>
> DLS, likewise, provides fundamental information on **particle size**, oligomerisation, aggregation state and purity of the samples. It is quick, cheap, easy and versatile as far as the range of concentration and formulation buffers

are concerned. It is therefore advisable that DLS measurements are performed in all cases.

Inaccurate **concentration estimation** can often be an issue as was observed in a few of the examined case studies, especially those included in the confidential Vol. 2. Concentration should therefore be checked and standardised for all samples (different batches of a protein), ideally as received on arrival, by UV-Vis spectroscopy. In the case that the excipients of the used formulation present high absorbance, resulting in low quality far-UV CD data, a diluted sample of each batch should be used for measuring the UV-Vis absorbance and estimating the concentration prior to all experiments.

For investigation of the **secondary structure**, either CD or FT-IR data should be acquired, depending on the protein concentration in the provided preparation. Namely, CD is recommended for relatively low concentrations (ideally 0.1–1 mg/mL and an upper extreme of 3 mg/mL), while FT-IR performs better at higher concentrations (ideally ≥10 mg/mL and a lower end at 1 mg/mL). Selecting between the two techniques should also be determined by the structure of the examined protein. As already seen in Chapter 4, certain protein structures can result in spectra that are indistinguishable from the spectrum of a completely different structure. Such was the case of protein A, where similarity of its unusual CD spectrum (of β_{II} type) with the typical random coil spectrum could raise doubts, while this issue could have probably been clarified by the position of the FT-IR amide I band if an FT-IR spectrum had been acquired. Vice versa, in the case of α-helix or random coil structure, which present overlapping amide I bands in FT-IR, CD can be used instead as the CD spectra for α-helix and random coil are fairly distinct. In principle, either CD or FT-IR should be sufficient for probing the secondary structure of a protein, while when in doubt both techniques can be used as complementary to each other.

For the **tertiary structure**, near-UV CD is probably sufficient for comparability purposes and can be easily recorded as an additional spectrum when far-UV CD data are recorded. However, in cases where major differences are apparent in the near-UV CD spectra or where changes in the

conformation of the protein are suspected, fluorescence is a useful tool, as shift in the maximum would directly address the question of any conformational change present. Fluorescence data would have probably been useful as an additional experiment in the case of the monoclonal antibody study (Section 4.7), where one of the samples (sample 4) appeared to be significantly different to the remaining three, and the reason of this variation was not clear. In addition, fluorescence is undoubtedly an advantageous technique for molecules with only a few aromatic amino acids (especially tyrosines and tryptophans) or even one tryptophan, as in such cases the emission spectrum will provide very specific information, or in cases where aromatic amino acids are located in regions that are important for binding and activity of the molecule.

For investigation of the **higher-order structure** as a whole of a protein molecule, namely secondary, tertiary and quaternary structure, the use of 1D ^{1}H-NMR spectroscopy seems to be promising. Acquisition of NMR spectra should be considered when dealing with high concentrations ($\geq$200 μM), as the provided data can offer complementary information to the rest of the techniques. NMR can contribute to quantitation of the observed differences when used in combination with statistical analysis. However, it is probably not sufficient yet as a stand-alone technique for this purpose, and further experience and data are needed in order to define the amount of variation among samples and the significance of such differences.

Overall, MS, DLS and far-UV CD or FT-IR (depending on protein structure- and preparation/concentration-specific experiment planning) are the most important techniques for assessing batch-to-batch or biosimilar-to-innovator similarity of biotherapeutic molecules. The rest of the techniques can provide additional and supporting information according to the manufacturer's requirements, in order to point out differences or confirm similarity among samples, and also clarify any unclear results. Concentrations should be checked for all samples upon arrival and prior to all experiments, either for the samples as received or following dilution. When dilution or buffer exchange is unavoidable, it should ideally be confirmed that the protein conformation does not change upon dilution (with an FT-IR or far-/near-UV CD experiment for example). Coupling with temperature provides additional information

on thermal stability evaluation and shows which samples are more prone to conformational changes under the stress of high temperature.

Future Aspects and Improvements

Even though exposing a protein to forced stress (such as high temperature) provides a valid comparability study on the susceptibility of each of the examined samples to conformational changes, it does not necessarily reflect changes that might occur to the molecule under the everyday-life stress conditions that a biopharmaceutical product can undergo during shipping, handling and storage processes. Long-term stability studies under relatively moderate stress and real-time storage conditions such as exposure to: direct sunlight, air (*e.g.* using a permeable vial), temperatures up to 45–60 °C (useful for countries with high environmental temperatures), high humidity, shaking, freeze-thaw cycles, long shelf-storage (*e.g.* ≥2 years), will probably be more indicative of the actual changes that might take place and the resulting degradation products. Temperature ramps that do not go as high as 75 °C or 100 °C will help in probing whether stability and folding are compromised at lower temperatures (*e.g.* 50–60 °C), if the protein returns to its native fold when cooled down to 20 °C and which is the upper temperature that represents the point of no return.

Finally, further NMR work in the direction of studying the glycan patterns following deglycosylation of a protein is probably an interesting complement to MS data, while statistical analysis of consistent NMR data and quantitation of the observed variation is an attractive addition to the rest of the analytical techniques.

Appendix A

Mass Spectrometry

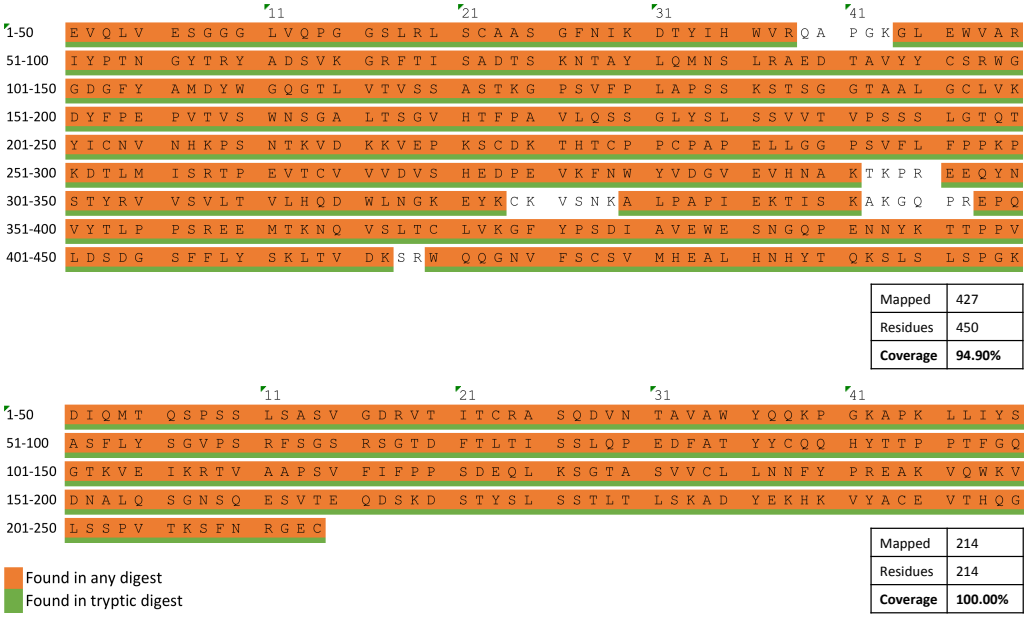


Figure A1: Overall coverage map for the trastuzumab sequence after tryptic digestion. The identified residues for the heavy (top) and the light (bottom) chain are marked in orange, and the missing peptides appear in white. Coverage was 94.90% for the heavy chain and 100.00% for the light chain.

Table A1: Peptide mapping table resulting from digestion with trypsin for the native antibody (sample mAb (S)). Calc. MW is the monoisotopic mass calculated from the m/z mass and the charge state. Exp. MW is the expected monoisotopic mass of the cleaved peptide.

RT	m/z	Charge state	Calc. MW (Da)	Chain	Possible Assignment	Exp. MW (Da)	Comments
10.70	939.94	2	1877.88	LC	1–18	1877.88	
8.13	947.96	2	1893.92	LC	1–18	1893.88	Oxidised
4.72	750.38	1	749.38	LC	19–24	749.37	
9.91	1161.55	1	1160.55	LC	25–35	1160.55	Cleaved at W
1.43	572.83	2	1143.66	LC	36–45	1143.64	Missed cleavage
16.96	886.98	2	1771.96	LC	46–61	1771.95	
0.55	553.28	1	552.28	LC	62–66	552.27	
19.42	1396.97	3	4187.91	LC	67–103	4187.89	
2.01	488.31	1	487.31	LC	104–107	487.30	
2.33	545.34	1	544.34	LC	105–108	544.33	Missed cleavage
16.69	973.52	2	1945.04	LC	109–126	1945.02	
13.99	752.39	3	2254.17	LC	126–145	2254.14	Missed cleavage
18.40	899.94	2	1797.88	LC	127–142	1797.87	
3.11	560.32	1	559.32	LC	146–149	559.31	
6.43	1068.49	2	2134.98	LC	150–169	2134.96	
12.04	751.88	2	1501.76	LC	170–183	1501.75	
0.47	625.28	1	624.28	LC	184–188	624.28	
8.14	714.68	3	2141.04	LC	189–207	2141.06	Missed cleavage
9.55	938.96	2	1875.92	LC	191–207	1875.90	
0.72	523.27	1	522.27	LC	208–211	522.26	
3.01	870.33	1	869.33	LC	208–214	869.33	Missed cleavage
12.55	941.51	2	1881.02	HC	1–19	1881.00	
9.52	584.78	2	1167.56	HC	20–30	1167.56	
15.52	661.98	3	1982.94	HC	20–36	1982.92	Missed cleavage, cleaved at W
14.63	747.04	3	2238.12	HC	20–38	2238.09	Missed cleavage
9.45	545.28	2	1088.56	HC	31–38	1088.54	
14.16	755.43	1	754.43	HC	37–43	754.44	Tentative
8.57	830.45	1	829.45	HC	44–50	829.44	
5.92	1084.54	1	1083.54	HC	51–59	1083.53	
6.11	1085.54	1	1084.54	HC	51–59	1084.53	Deamidated
2.04	682.34	1	681.34	HC	60–65	681.33	
6.07	591.80	2	1181.60	HC	66–76	1181.60	Missed cleavage
6.69	969.49	1	968.49	HC	68–76	968.48	
9.87	655.83	2	1309.66	HC	77–87	1309.64	
6.49	668.28	2	1334.56	HC	88–98	1334.54	
18.58	1392.62	2	2783.24	HC	99–124	2783.25	
11.04	593.83	2	1185.66	HC	125–136	1185.64	
10.21	661.84	2	1321.68	HC	137–150	1321.65	
22.47	1343.67	5	6713.35	HC	151–213	6713.29	Missed cleavage
12.33	726.86	4	2903.44	HC	192–217	2903.44	Missed cleavage
2.99	361.21	1	360.21	HC	214–216	360.20	Weak
13.06	472.28	1	471.28	HC	218–221	471.27	
15.89	844.04	3	2529.12	HC	222–244	2529.09	Missed cleavage, cleaved at F
16.52	1019.96	2	2037.92	HC	226–244	2037.92	Cleaved at F
6.07	826.52	1	825.52	HC	245–251	825.51	Missed cleavage
6.96	835.44	1	834.44	HC	252–258	834.43	
5.89	851.42	1	850.42	HC	252–258	850.43	Oxidised
13.06	714.01	3	2139.03	HC	259–277	2139.00	
16.04	763.03	3	2286.09	HC	259–278	2286.07	Missed cleavage, cleaved at F
11.79	839.40	2	1676.80	HC	278–291	1676.79	
4.31	1317.54	2	2633.08	HC	296–304	2633.00	G0F 1444.50
4.31	1398.55	2	2795.10	HC	296–304	2795.08	G1F 1606.58
4.31	1479.57	2	2957.14	HC	296–304	2957.13	G2F 1768.63
17.37	904.50	2	1807.00	HC	305–320	1807.00	
17.75	904.98	2	1807.96	HC	305–320	1808.00	Deamidated
16.50	743.41	3	2227.23	HC	305–323	2227.20	Missed cleavage
16.74	743.73	3	2228.19	HC	305–323	2228.20	Missed cleavage deamidated
6.09	578.79	2	1155.58	HC	321–329	1155.56	Tentative, missed cleavage
7.30	838.50	1	837.50	HC	330–337	837.50	
0.36	448.27	1	447.27	HC	338–341	447.27	
12.82	656.37	1	655.37	HC	342–347	655.38	Tentative, missed cleavage
16.72	884.45	2	1766.90	HC	342–357	1766.93	Tentative, missed cleavage
10.21	643.82	2	1285.64	HC	348–358	1285.67	
0.47	637.29	1	636.29	HC	359–363	636.28	Very weak
11.43	581.81	2	1161.62	HC	364–373	1161.61	
14.87	1272.56	2	2543.12	HC	374–395	2543.12	
15.11	1273.06	2	2544.12	HC	374–395	2544.12	Deamidated
15.70	937.47	2	1872.94	HC	396–412	1872.91	
2.99	575.34	1	574.34	HC	413–417	574.33	
12.82	1201.24	3	3600.72	HC	413–442	3600.70	Tentative, missed cleavage
12.69	701.31	4	2801.24	HC	420–442	2801.24	
9.55	788.46	1	787.46	HC	443–450	787.44	Very weak
7.55	660.36	1	659.36	HC	443–450	659.35	K missing

Table A2: Peptide mapping table resulting from digestion with trypsin for the native antibody (sample mAb (LR)). Calc. MW is the monoisotopic mass calculated from the m/z mass and the charge state. Exp. MW is the expected monoisotopic mass of the cleaved peptide.

RT	m/z	Charge state	Calc. MW (Da)	Chain	Possible Assignment	Exp. MW (Da)	Comments
10.70	939.94	2	1877.88	LC	1–18	1877.88	
8.14	947.94	2	1893.88	LC	1–18	1893.88	Oxidised
4.72	750.38	1	749.38	LC	19–24	749.37	
9.01	763.07	3	2286.21	LC	25–45	2286.18	Missed cleavage
16.96	886.98	2	1771.96	LC	46–61	1771.95	
0.55	553.27	1	552.27	LC	62–66	552.27	
19.43	1396.98	3	4187.94	LC	67–103	4187.89	
2.01	488.31	1	487.31	LC	104–107	487.30	
2.31	644.41	1	643.41	LC	104–108	643.40	Missed cleavage
16.72	973.52	2	1945.04	LC	109–126	1945.02	
14.00	752.39	3	2254.17	LC	126–145	2254.14	Missed cleavage
18.40	899.94	2	1797.88	LC	127–142	1797.87	
3.96	888.48	1	887.48	LC	143–149	887.49	Missed cleavage
3.12	560.32	1	559.32	LC	146–149	559.31	
6.43	1068.49	2	2134.98	LC	150–169	2134.96	
12.04	751.88	2	1501.76	LC	170–183	1501.75	
0.48	625.29	1	624.29	LC	184–188	624.28	
8.14	714.69	3	2141.07	LC	189–207	2141.06	Missed cleavage
9.55	938.96	2	1875.92	LC	191–207	1875.90	
0.70	523.27	1	522.27	LC	208–211	522.26	
3.01	870.35	1	869.35	LC	208–214	869.33	Missed cleavage
12.55	941.50	2	1881.00	HC	1–19	1881.00	
9.52	584.78	2	1167.56	HC	20–30	1167.56	
15.55	661.98	3	1982.94	HC	20–36	1982.92	Missed cleavage, cleaved at W
14.65	747.04	3	2238.12	HC	20–38	2238.09	Missed cleavage
9.43	545.28	2	1088.56	HC	31–38	1088.54	
14.16	755.44	1	754.44	HC	37–43	754.44	Tentative
0.33	500.28	1	499.28	HC	39–43	499.28	Tentative
8.57	830.45	1	829.45	HC	44–50	829.44	
5.93	1084.54	1	1083.54	HC	51–59	1083.53	
6.09	1085.53	1	1084.53	HC	51–59	1084.53	Deamidated
2.04	682.34	1	681.34	HC	60–65	681.33	
6.07	591.80	2	1181.60	HC	66–76	1181.60	Missed cleavage
6.70	969.49	1	968.49	HC	68–76	968.48	
9.89	655.83	2	1309.66	HC	77–87	1309.64	
6.49	1335.56	1	1334.56	HC	88–98	1334.54	
18.57	1392.63	2	2783.26	HC	99–124	2783.25	
11.04	593.83	2	1185.66	HC	125–136	1185.64	
10.21	661.84	2	1321.68	HC	137–150	1321.65	
22.49	1343.65	5	6713.25	HC	151–213	6713.29	Missed cleavage
12.35	968.82	3	2903.46	HC	192–217	2903.44	Missed cleavage
9.33	765.87	2	1529.74	HC	213–225	1529.78	Missed cleavage, weak
13.07	472.28	1	471.28	HC	218–221	471.27	Weak
15.91	1265.55	2	2529.10	HC	222–244	2529.09	Missed cleavage, cleaved at F
16.55	1019.96	2	2037.92	HC	226–244	2037.92	Cleaved at F
5.91	826.52	1	825.52	HC	245–251	825.51	Missed cleavage
5.89	566.37	1	565.37	HC	247–251	565.36	Missed cleavage
6.96	835.43	1	834.43	HC	252–258	834.43	
5.89	851.42	1	850.42	HC	252–258	850.43	Oxidised
13.07	714.01	3	2139.03	HC	259–277	2139.00	
16.00	763.03	3	2286.09	HC	259–278	2286.07	Missed cleavage, cleaved at F
11.79	839.40	2	1676.80	HC	278–291	1676.79	
10.37	595.31	2	1188.62	HC	296–304	1188.50	Very weak
4.31	1317.51	2	2633.02	HC	296–304	2633.00	G0F 1444.50
4.31	1398.55	2	2795.10	HC	296–304	2795.08	G1F 1606.58
4.31	1479.58	2	2957.16	HC	296–304	2957.13	G2F 1768.63
17.37	904.51	2	1807.02	HC	305–320	1807.00	
17.74	905.00	2	1808.00	HC	305–320	1808.00	Deamidated
16.55	743.40	3	2227.20	HC	305–323	2227.20	Missed cleavage
16.75	743.74	3	2228.22	HC	305–323	2228.20	Missed cleavage deamidated
6.11	1156.58	1	1155.58	HC	321–329	1155.56	Tentative, missed cleavage
0.36	736.37	1	735.37	HC	324–329	735.36	Missed cleavage, weak
7.31	838.50	1	837.50	HC	330–337	837.50	
0.36	448.28	1	447.28	HC	338–341	447.27	
12.86	656.37	1	655.37	HC	342–347	655.38	Tentative, missed cleavage
16.67	884.45	2	1766.90	HC	342–357	1766.93	Tentative, missed cleavage
10.21	643.82	2	1285.64	HC	348–358	1285.67	
0.50	637.28	1	636.28	HC	359–363	636.28	Very weak
11.43	581.82	2	1161.64	HC	364–373	1161.61	
14.87	1272.57	2	2543.14	HC	374–395	2543.12	
15.06	1273.07	2	2544.14	HC	374–395	2544.12	Deamidated
15.69	937.47	2	1872.94	HC	396–412	1872.91	
2.99	575.34	1	574.34	HC	413–417	574.33	
12.84	901.19	4	3600.76	HC	413–442	3600.70	Tentative, missed cleavage
12.69	934.76	3	2801.28	HC	420–442	2801.24	
8.14	788.45	1	787.45	HC	443–450	787.44	

Table A3: Peptide mapping table resulting from digestion with trypsin for the antibody drug conjugate produced with the solution method (sample ADC (S)). Calc. MW is the monoisotopic mass calculated from the m/z mass and the charge state. Exp. MW is the expected monoisotopic mass of the cleaved peptide.

RT	m/z	Charge state	Calc. MW (Da)	Chain	Possible Assignment	Exp. MW (Da)	Comments
10.74	939.95	2	1877.90	LC	1–18	1877.88	
8.18	947.94	2	1893.88	LC	1–18	1893.88	Oxidised
4.75	750.38	1	749.38	LC	19–24	749.37	
9.94	1161.54	1	1160.54	LC	25–35	1160.55	Cleaved at W
1.50	572.83	2	1143.66	LC	36–45	1143.64	Missed cleavage
16.96	886.98	2	1771.96	LC	46–61	1771.95	
0.55	553.27	1	552.27	LC	62–66	552.27	
19.47	1396.97	3	4187.91	LC	67–103	4187.89	
2.04	488.31	1	487.31	LC	104–107	487.30	
9.30	947.54	1	946.54	LC	108–116	946.52	Cleaved at F
16.64	973.52	2	1945.04	LC	109–126	1945.02	
9.04	457.24	2	912.48	LC	119–126	912.46	
14.00	752.39	3	2254.17	LC	126–145	2254.14	Missed cleavage
18.44	899.95	2	1797.90	LC	127–142	1797.87	
3.14	560.32	1	559.32	LC	146–149	559.31	
6.45	1068.49	2	2134.98	LC	150–169	2134.96	
12.06	751.88	2	1501.76	LC	170–183	1501.75	
0.47	625.28	1	624.28	LC	184–188	624.28	
8.19	714.69	3	2141.07	LC	189–207	2141.06	Missed cleavage
9.58	938.95	2	1875.90	LC	191–207	1875.90	
0.70	523.26	1	522.26	LC	208–211	522.26	
3.04	870.34	1	869.34	LC	208–214	869.33	Missed cleavage, unconjugated
23.47	1064.56	2	2127.12	LC	208–214	2127.12	Missed cleavage, drug peptide
25.09	812.44	2	1622.88	LC	212–214	1622.87	Drug peptide
12.57	941.50	2	1881.00	HC	1–19	1881.00	
9.57	584.79	2	1167.58	HC	20–30	1167.56	
15.55	661.98	3	1982.94	HC	20–36	1982.92	Missed cleavage, cleaved at W
14.69	747.04	3	2238.12	HC	20–38	2238.09	Missed cleavage
14.16	755.42	1	754.42	HC	37–43	754.44	Tentative
8.62	830.45	1	829.45	HC	44–50	829.44	
5.94	1084.55	1	1083.55	HC	51–59	1083.53	
6.13	1085.56	1	1084.56	HC	51–59	1084.53	Deamidated
2.09	682.35	1	681.35	HC	60–65	681.33	
6.13	591.81	2	1181.62	HC	66–76	1181.60	Missed cleavage
6.74	969.48	1	968.48	HC	68–76	968.48	
9.96	655.83	2	1309.66	HC	77–87	1309.64	
6.50	668.29	2	1334.58	HC	88–98	1334.54	
18.57	1392.64	2	2783.28	HC	99–124	2783.25	
11.07	593.83	2	1185.66	HC	125–136	1185.64	
10.26	661.83	2	1321.66	HC	137–150	1321.65	
21.95	1176.22	3	3525.66	HC	151–183	3525.71	Cleaved at F
22.50	1343.66	5	6713.30	HC	151–213	6713.29	Missed cleavage
12.88	887.96	4	3547.84	HC	184–216	3547.78	Missed cleavage
12.38	726.86	4	2903.44	HC	192–217	2903.44	Missed cleavage
13.11	472.28	1	471.28	HC	218–221	471.27	
22.84	884.48	2	1766.96	HC	222–225	1766.96	Drug peptide
15.98	844.04	3	2529.12	HC	222–244	2529.09	Missed cleavage, cleaved at F, unconjugated
25.65	1263.29	3	3786.87	HC	222–244	3786.87	Missed cleavage, drug peptide
16.57	1019.96	2	2037.92	HC	226–244	2037.92	Cleaved at F
5.94	826.52	1	825.52	HC	245–251	825.51	Missed cleavage
6.99	835.43	1	834.43	HC	252–258	834.43	
5.93	851.45	1	850.45	HC	252–258	850.43	Oxidised
13.11	714.01	3	2139.03	HC	259–277	2139.00	
16.08	763.03	3	2286.09	HC	259–278	2286.07	Missed cleavage, cleaved at F
11.84	839.40	2	1676.80	HC	278–291	1676.79	
10.42	595.30	2	1188.60	HC	296–304	1188.50	Very weak
4.33	1317.51	2	2633.02	HC	296–304	2633.00	G0F 1444.50
4.33	1398.55	2	2795.10	HC	296–304	2795.08	G1F 1606.58
4.33	1479.58	2	2957.16	HC	296–304	2957.13	G2F 1768.63
17.42	904.51	2	1807.02	HC	305–320	1807.00	
17.77	905.00	2	1808.00	HC	305–320	1808.00	Deamidated
16.57	743.41	3	2227.23	HC	305–323	2227.20	Missed cleavage
16.81	743.73	3	2228.19	HC	305–323	2228.20	Missed cleavage, deamidated
6.16	578.78	2	1155.56	HC	321–329	1155.56	Tentative, missed cleavage
7.35	838.50	1	837.50	HC	330–337	837.50	
0.36	448.28	1	447.28	HC	338–341	447.27	
16.69	884.44	2	1766.88	HC	342–357	1766.93	Tentative, missed cleavage
10.26	643.83	2	1285.66	HC	348–358	1285.67	
0.50	637.29	1	636.29	HC	359–363	636.28	Very weak
11.47	581.81	2	1161.62	HC	364–373	1161.61	
14.89	1272.56	2	2543.12	HC	374–395	2543.12	
15.06	1273.08	2	2544.16	HC	374–395	2544.12	Deamidated
15.69	937.47	2	1872.94	HC	396–412	1872.91	
2.99	575.34	1	574.34	HC	413–417	574.33	
12.86	901.18	4	3600.72	HC	413–442	3600.70	Tentative, missed cleavage
12.72	934.75	3	2801.25	HC	420–442	2801.24	
8.19	788.45	1	787.45	HC	443–450	787.44	Very weak
7.58	660.36	1	659.36	HC	443–450	659.35	K missing

Table A4: Peptide mapping table resulting from digestion with trypsin for the antibody drug conjugate produced with the lock-release method (sample ADC (LR)). Calc. MW is the monoisotopic mass calculated from the m/z mass and the charge state. Exp. MW is the expected monoisotopic mass of the cleaved peptide.

RT	m/z	Charge state	Calc. MW (Da)	Chain	Possible Assignment	Exp. MW (Da)	Comments
10.72	939.95	2	1877.90	LC	1–18	1877.88	
8.16	947.94	2	1893.88	LC	1–18	1893.88	Oxidised
4.74	750.38	1	749.38	LC	19–24	749.37	
9.93	1161.55	1	1160.55	LC	25–35	1160.55	Cleaved at W
1.47	572.83	2	1143.66	LC	36–45	1143.64	Missed cleavage
16.98	886.98	2	1771.96	LC	46–61	1771.95	
0.55	553.27	1	552.27	LC	62–66	552.27	
19.43	1396.97	3	4187.91	LC	67–103	4187.89	
21.93	1183.58	3	3547.74	LC	78–107	3547.66	Missed cleavage
2.04	488.31	1	487.31	LC	104–107	487.30	
2.31	545.34	1	544.34	LC	105–108	544.33	Missed cleavage
16.74	973.52	2	1945.04	LC	109–126	1945.02	
13.99	752.39	3	2254.17	LC	126–145	2254.14	Missed cleavage
18.42	899.94	2	1797.88	LC	127–142	1797.87	
10.72	347.17	1	346.17	LC	143–145	346.19	Very weak
3.14	560.32	1	559.32	LC	146–149	559.31	
6.43	1068.48	2	2134.96	LC	150–169	2134.96	
12.06	751.88	2	1501.76	LC	170–183	1501.75	
0.47	625.28	1	624.28	LC	184–188	624.28	
8.16	714.69	3	2141.07	LC	189–207	2141.06	Missed cleavage
9.57	938.96	2	1875.92	LC	191–207	1875.90	
0.70	523.26	1	522.26	LC	208–211	522.26	
3.04	870.34	1	869.34	LC	208–214	869.33	Missed cleavage, unconjugated
23.45	1064.56	2	2127.12	LC	208–214	2127.12	Missed cleavage, drug peptide
25.11	812.43	2	1622.86	LC	212–214	1622.87	Drug peptide
12.55	941.50	2	1881.00	HC	1–19	1881.00	
9.55	584.78	2	1167.56	HC	20–30	1167.56	
15.57	661.98	3	1982.94	HC	20–36	1982.92	Missed cleavage, cleaved at W
14.67	747.04	3	2238.12	HC	20–38	2238.09	Missed cleavage
9.47	545.28	2	1088.56	HC	31–38	1088.54	
14.16	755.43	1	754.43	HC	37–43	754.44	Tentative
8.58	830.46	1	829.46	HC	44–50	829.44	
5.93	1084.54	1	1083.54	HC	51–59	1083.53	
6.13	1085.54	1	1084.54	HC	51–59	1084.53	Deamidated
2.04	682.34	1	681.34	HC	60–65	681.33	
6.09	591.80	2	1181.60	HC	66–76	1181.60	Missed cleavage
6.70	969.49	1	968.49	HC	68–76	968.48	
9.93	655.83	2	1309.66	HC	77–87	1309.64	
6.50	668.27	2	1334.54	HC	88–98	1334.54	
18.57	1392.63	2	2783.26	HC	99–124	2783.25	
11.04	593.83	2	1185.66	HC	125–136	1185.64	
10.23	661.84	2	1321.68	HC	137–150	1321.65	
22.49	1343.66	5	6713.30	HC	151–213	6713.29	Missed cleavage
12.35	726.87	4	2903.48	HC	192–217	2903.44	Missed cleavage
2.99	361.21	1	360.21	HC	214–216	360.20	Weak
13.09	472.28	1	471.28	HC	218–221	471.27	
22.79	884.49	2	1766.98	HC	222–225	1766.96	Drug peptide
15.96	844.03	3	2529.09	HC	222–244	2529.09	Missed cleavage, cleaved at F, unconjugated
25.67	1263.29	3	3786.87	HC	222–244	3786.87	Missed cleavage, drug peptide
16.55	1019.95	2	2037.90	HC	226–244	2037.92	Cleaved at F
6.07	826.51	1	825.51	HC	245–251	825.51	Missed cleavage
6.96	835.44	1	834.44	HC	252–258	834.43	
5.93	851.43	1	850.43	HC	252–258	850.43	Oxidised
13.09	714.01	3	2139.03	HC	259–277	2139.00	
16.04	763.03	3	2286.09	HC	259–278	2286.07	Missed cleavage, cleaved at F
11.82	839.40	2	1676.80	HC	278–291	1676.79	
4.33	1317.52	2	2633.04	HC	296–304	2633.00	G0F 1444.50
4.31	1398.54	2	2795.08	HC	296–304	2795.08	G1F 1606.58
4.33	1479.58	2	2957.16	HC	296–304	2957.13	G2F 1768.63
17.40	904.51	2	1807.02	HC	305–320	1807.00	
17.77	905.01	2	1808.02	HC	305–320	1808.00	Deamidated
16.55	743.41	3	2227.23	HC	305–323	2227.20	Missed cleavage
16.77	743.73	3	2228.19	HC	305–323	2228.20	Missed cleavage, deamidated
7.33	838.50	1	837.50	HC	330–337	837.50	
0.36	448.28	1	447.28	HC	338–341	447.27	
16.60	884.45	2	1766.90	HC	342–357	1766.93	Tentative, missed cleavage
10.21	643.82	2	1285.64	HC	348–358	1285.67	
0.47	637.28	1	636.28	HC	359–363	636.28	
11.45	581.82	2	1161.64	HC	364–373	1161.61	
14.87	1272.56	2	2543.12	HC	374–395	2543.12	
15.06	1273.06	2	2544.12	HC	374–395	2544.12	Deamidated
15.69	937.47	2	1872.94	HC	396–412	1872.91	
3.01	575.34	1	574.34	HC	413–417	574.33	
12.86	901.17	4	3600.68	HC	413–442	3600.70	Weak
12.72	701.32	4	2801.28	HC	420–442	2801.24	
8.16	788.44	1	787.44	HC	443–450	787.44	
7.57	660.36	1	659.36	HC	443–450	659.35	K missing

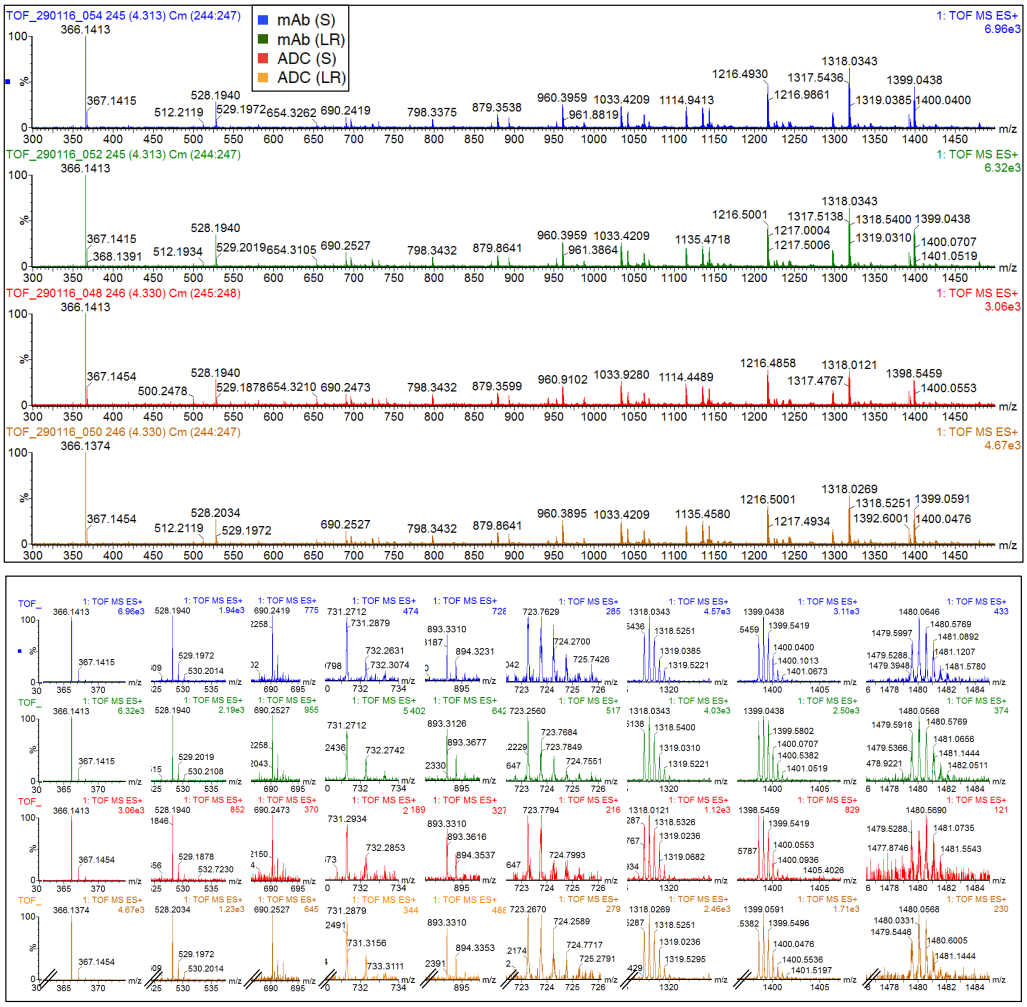


Figure A2: Comparison of the MS spectra of the glycosylated peptide HC296–304 for all four samples (top). The same peaks corresponding to the glycosylated peptide and glycan fragments (as described in Figure 3.2) were found in all cases as shown in the zoom-in image (bottom).

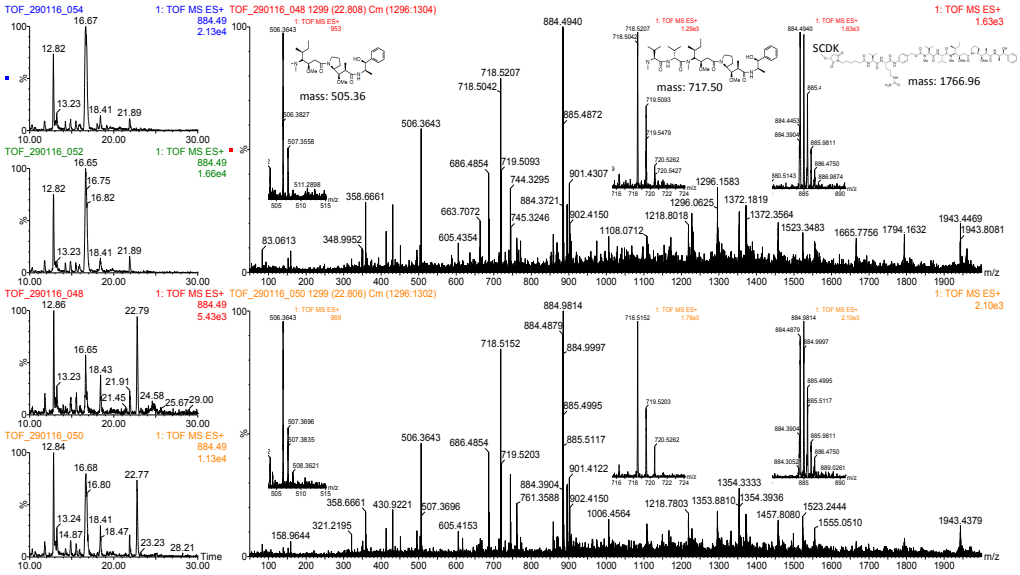


Figure A3: Spectrum of the drug-conjugated peptide HC222–225 (RT 22.81 min) for the ADC (S) (top, red) and ADC (LR) (bottom, orange) samples. The drug molecule is attached to Cys (223) resulting in

m/z 884.48 for the $[M+2H]^{2+}$ state. Signature fragment ions from vcMMAE are present as $[M+H]^{+}$, with m/z 718.51 (MMAE molecule) and 506.36. The extracted chromatograms on the left show the absence of the peak at ~22.80 min for the mAb samples (blue, green) in contrast to the ADC samples (red, orange).

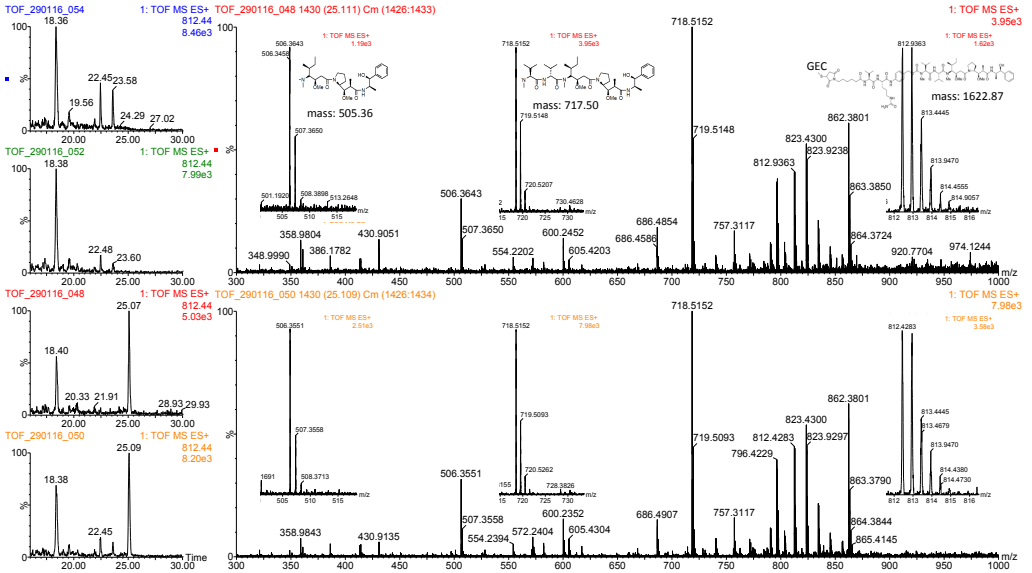


Figure A4: Spectrum of the drug-conjugated peptide LC212–214 (RT 25.11 min) for the ADC (S) (top, red) and ADC (LR) (bottom, orange) samples. The drug molecule is attached to Cys (214) resulting in m/z 812.44 for the $[M+2H]^{2+}$ state. Signature fragment ions from vcMMAE are present as $[M+H]^{+}$, with m/z 718.51 (MMAE molecule) and 506.36. The extracted chromatograms on the left show the absence of the peak at ~25.10 min for the mAb samples (blue, green) in contrast to the ADC samples (red, orange).

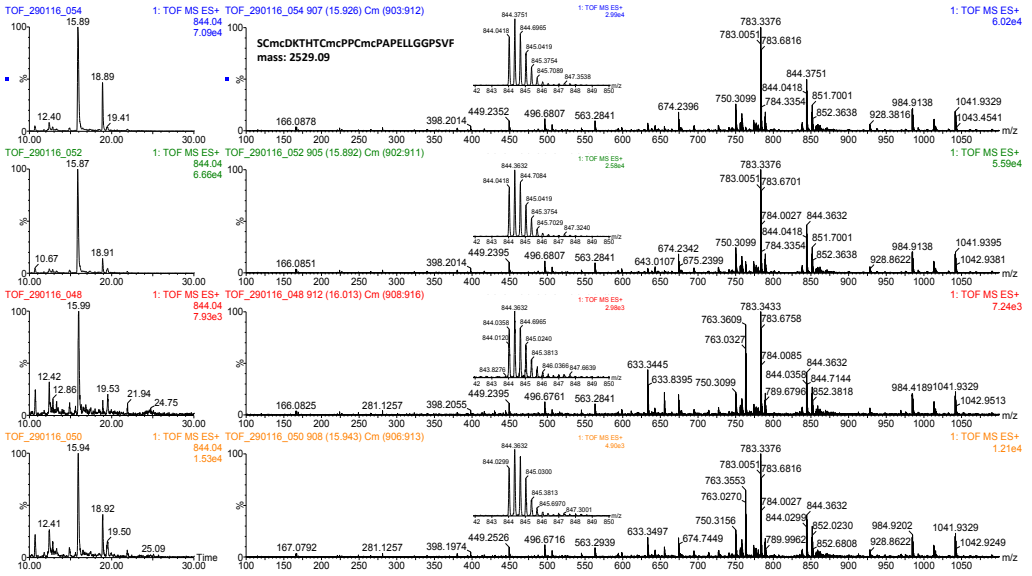


Figure A5: Comparison of the spectra of the non-conjugated peptide HC222–244 (RT ~16 min) for the four samples (mAb (S): blue, mAb (LR): green, ADC (S): red, ADC (LR): orange). All cysteines are carboxymethylated (Cmc) resulting in m/z 844.04 for the $[M+3H]^{3+}$ state. The extracted chromatograms on the left show the presence of the peak at ~16 min for all samples.

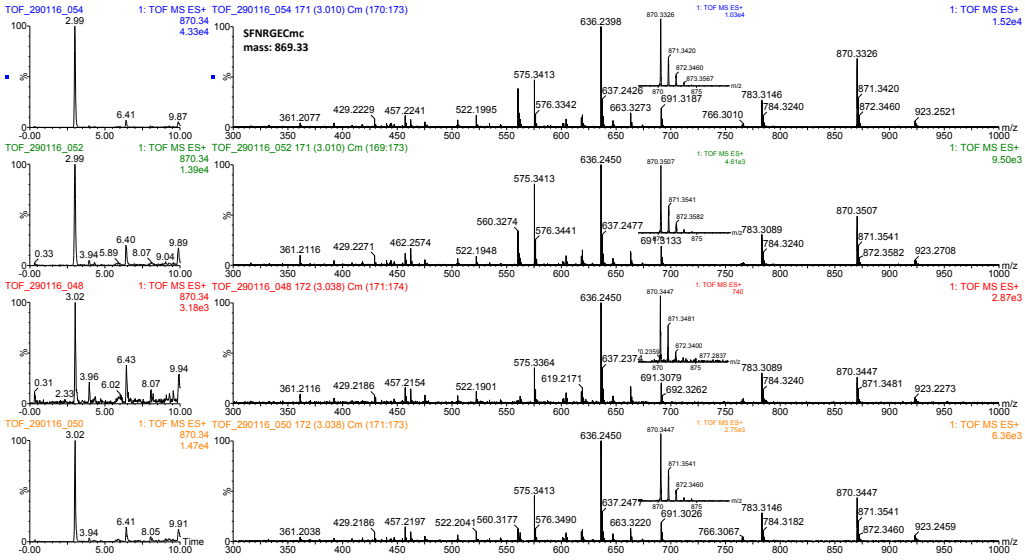


Figure A6: Comparison of the spectra of the non-conjugated peptide LC208–214 (RT ~3 min) for the four samples (mAb (S): blue, mAb (LR): green, ADC (S): red, ADC (LR): orange). Cys (214) is carboxymethylated (Cmc) resulting in m/z 870.34 for the $[M+H]^+$ state. The extracted chromatograms on the left show the presence of the peak at ~3 min in all samples.

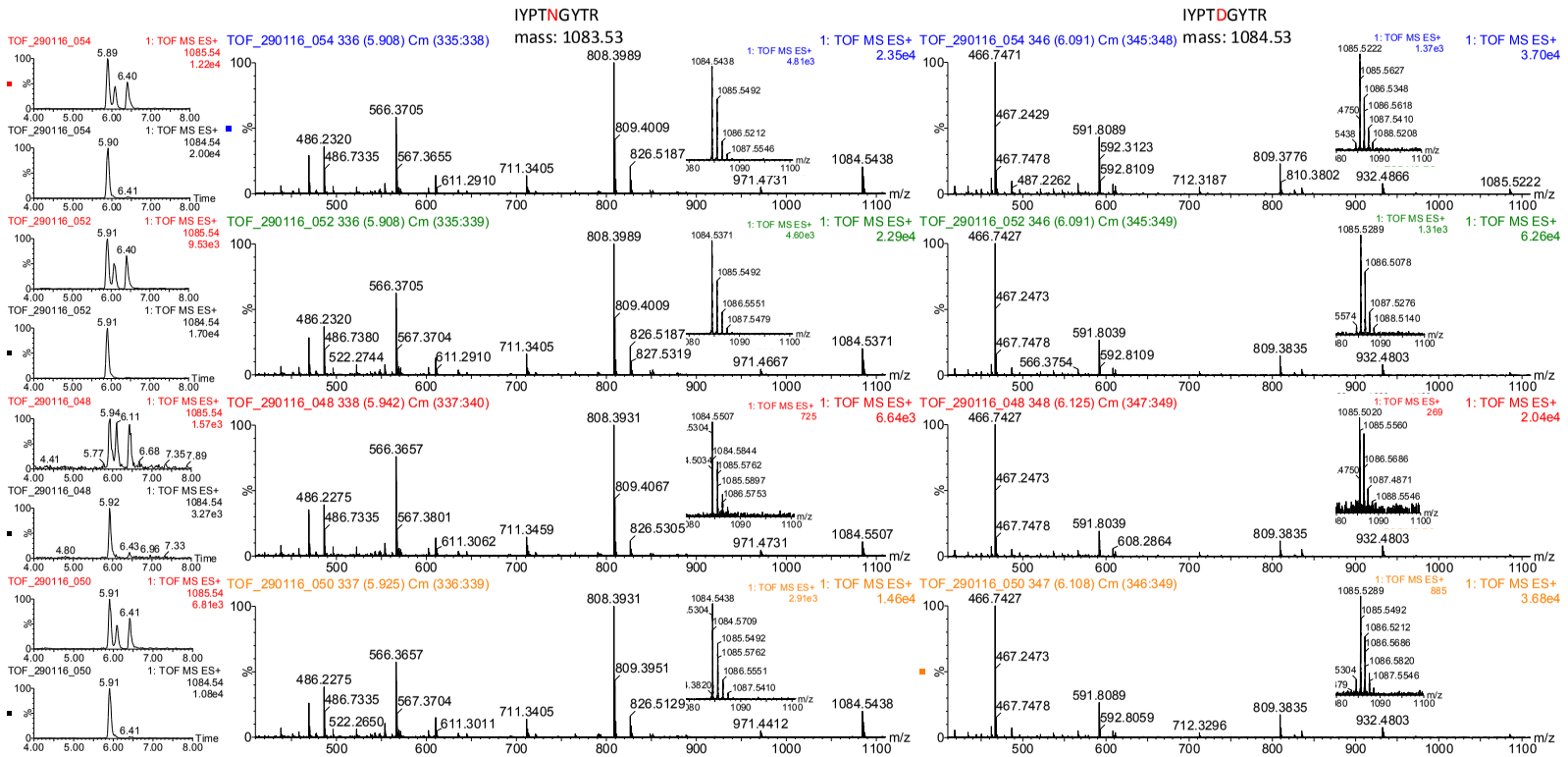


Figure A7: Comparison of the spectra of the native (left) and deamidated (right) peptide HC51–59 with RT ~5.9 min and ~6.1 min respectively are shown for the four samples (mAb (S): blue, mAb (LR): green, ADC (S): red, ADC (LR): orange). N55 is modified to aspartic acid (Asp) resulting in a mass increase of 1 Da and m/z 1085.53 for the $[M+H]^+$ state. The extracted chromatograms on the left correspond to the deamidated (red) and native (black) peptide for each sample. Except from the peak at ~5.9 min which is present in all cases, two additional peaks can be seen for the deamidated peptide at ~6.1 min and ~6.4 min probably corresponding to the Asp and isoAsp forms.

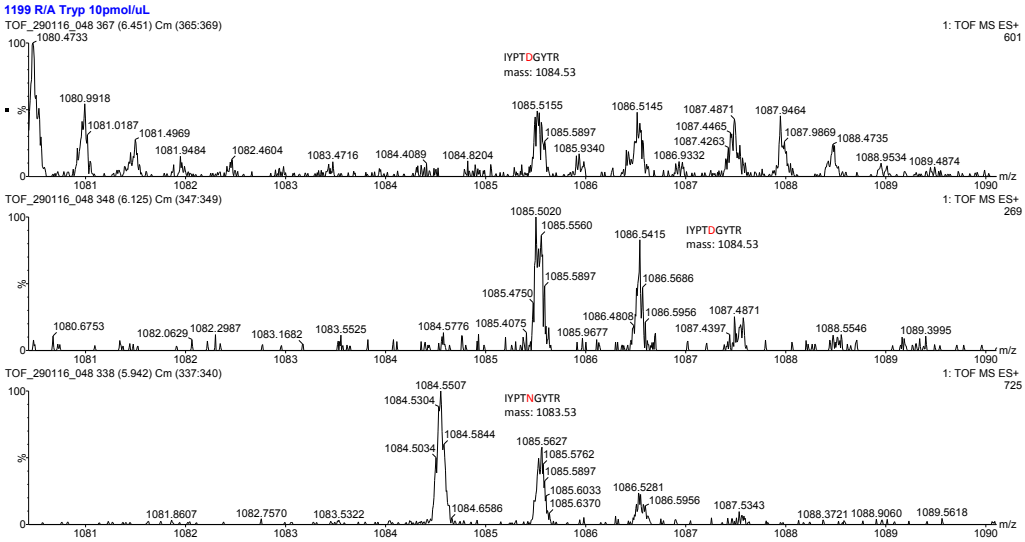


Figure A8: Spectra of the native (bottom) and deamidated (centre and top) peptide HC51–59 with RT 5.94 min, 6.12 min and 6.45 min respectively are shown for the ADC (S) sample. The two deamidated forms of N55 probably correspond to Asp and isoAsp formation. The weak intensities though, make it hard to differentiate between the two structures by looking at the fragment ions.

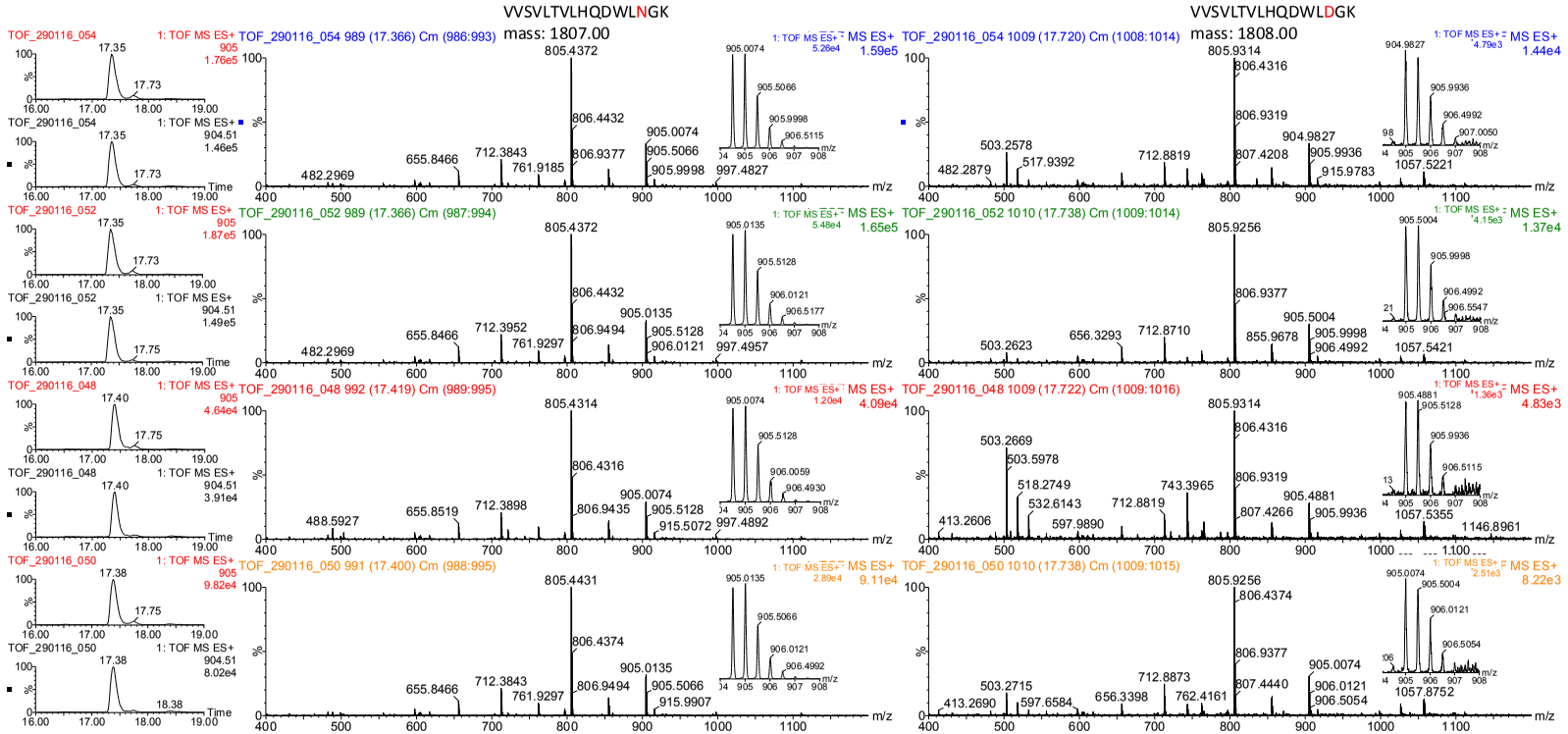


Figure A9: Comparison of the spectra of the native (left) and deamidated (right) peptide HC305–320 with RT ~17.4 min and ~17.7 min respectively are shown for the four samples (mAb (S): blue, mAb (LR): green, ADC (S): red, ADC (LR): orange). N318 is modified to aspartic acid (Asp) resulting in a mass increase of 1 Da and m/z 905.00 for the $[M+2H]^{2+}$ state. The extracted chromatograms on the left correspond to the deamidated (red) and native (black) peptide for each sample.

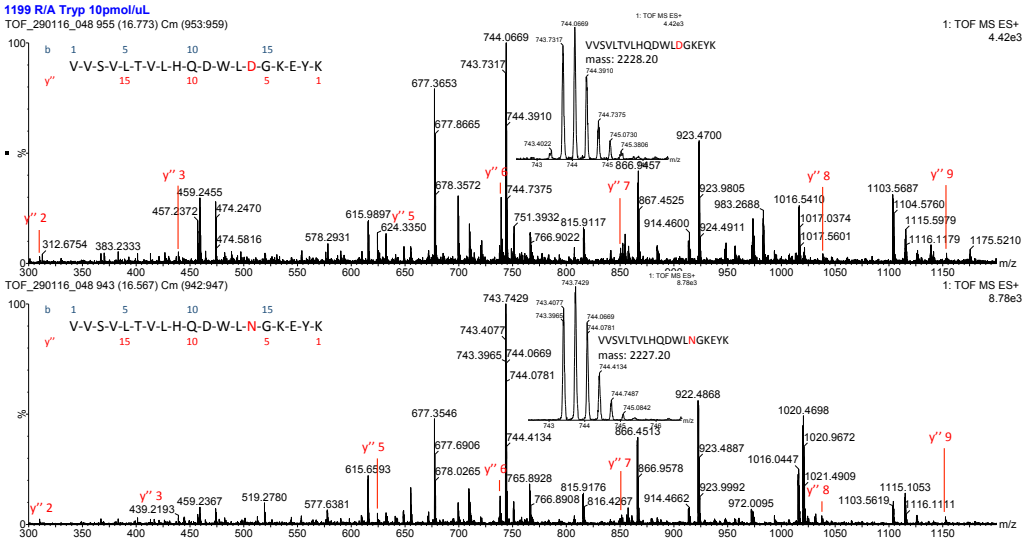


Figure A10: Spectra of the native (bottom) and deamidated (top) peptide HC305–323 with RT 16.57 min and 16.77 min respectively are shown for the ADC (S) sample. N318 is modified to aspartic acid resulting in a mass increase of 1 Da and m/z 743.74 for the $[M+3H]^{3+}$ state. The observed y'' fragment ions are assigned. A comparison of all four samples can be found in Figure A11.

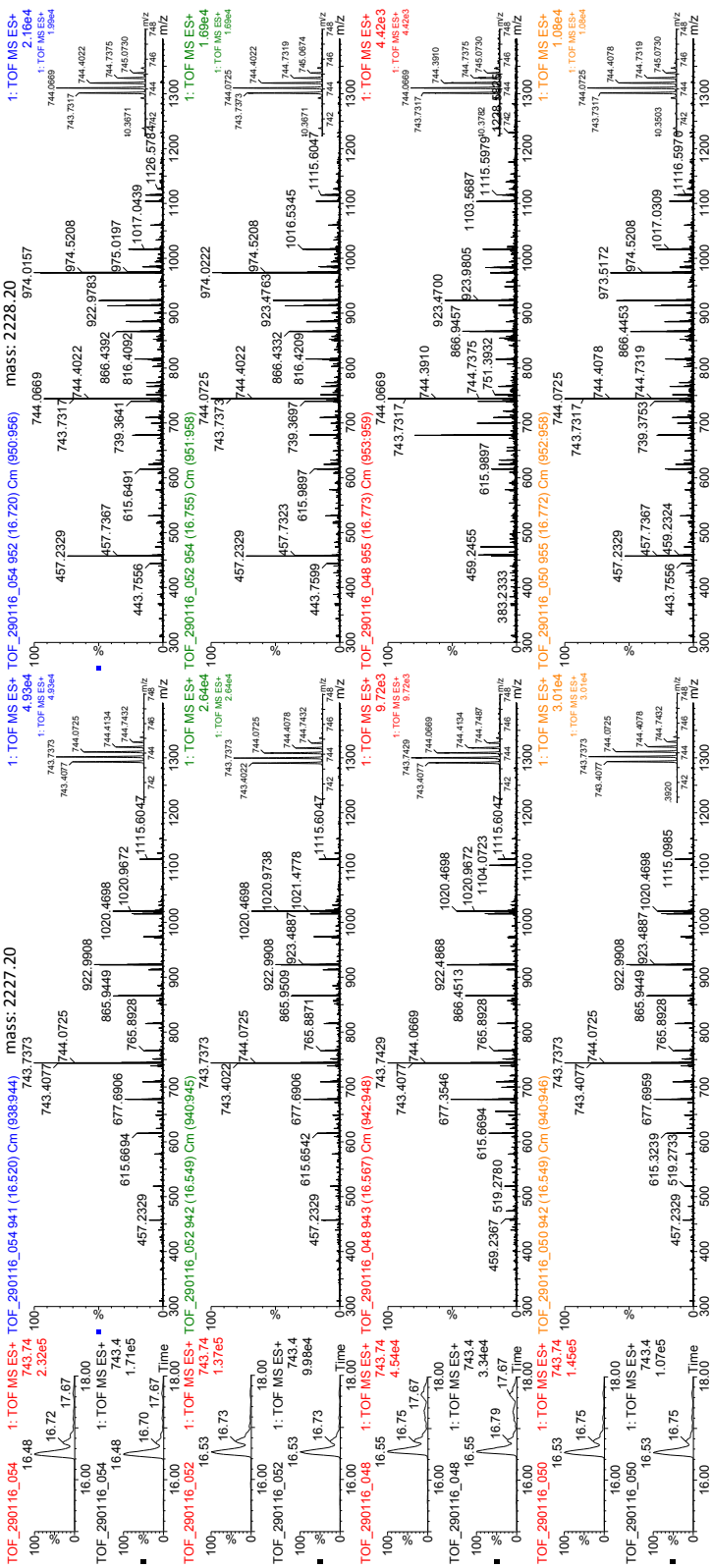


Figure A11: Comparison of the spectra of the native (left) and deamidated (right) peptide HC305–323 with RT ~16.5 min and ~16.7 min respectively are shown for the four samples (mAb (S): blue, mAb (LR): green, ADC (S): red, ADC (LR): orange). N318 is modified to aspartic acid (Asp) resulting in a mass increase of 1 Da and m/z 743.74 for the $[M+3H]^{3+}$ state. The extracted chromatograms on the left correspond to the deamidated (red) and native (black) peptide for each sample.

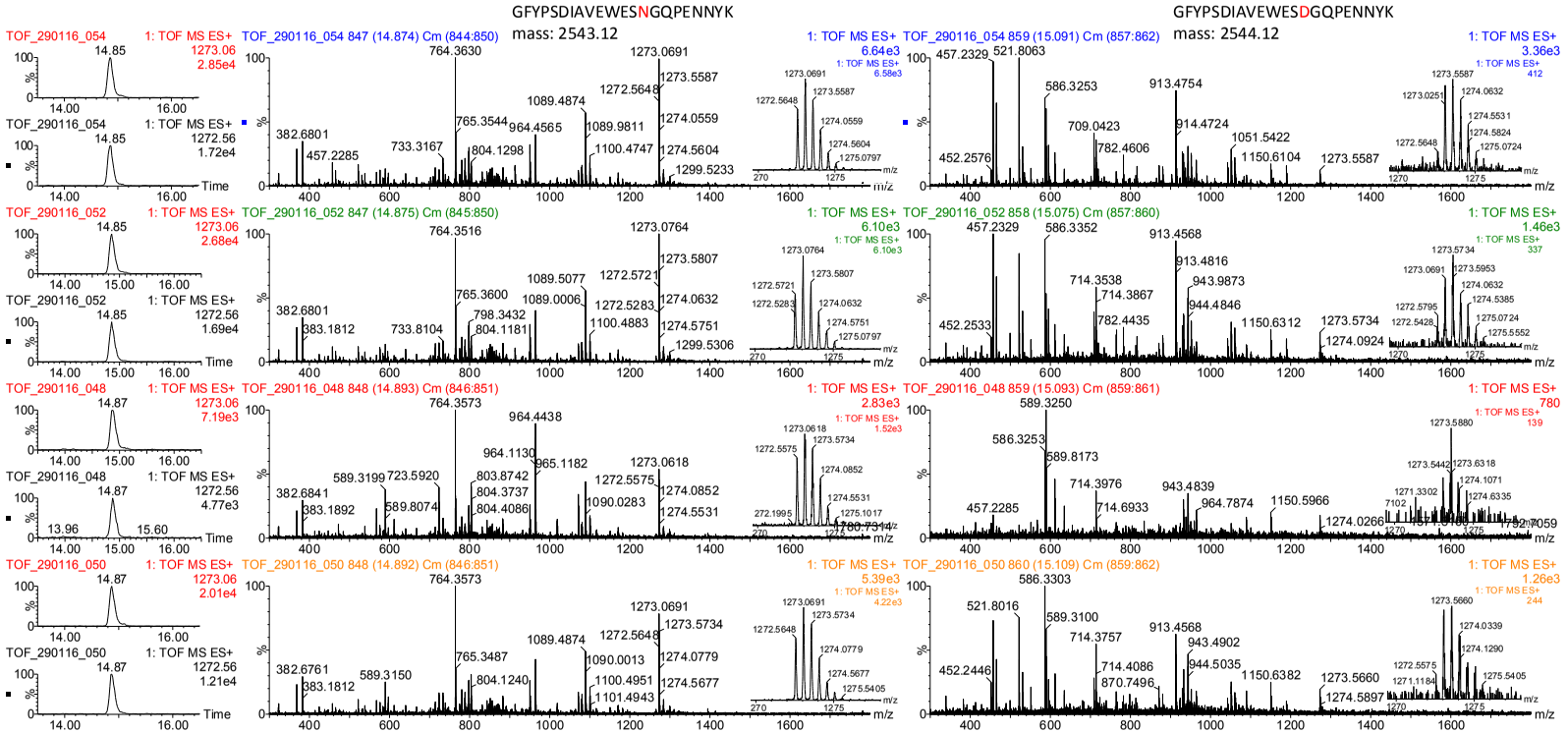


Figure A12: Comparison of the spectra of the native (left) and deamidated (right) peptide HC374–395 with RT ~14.8 min and ~15.1 min respectively are shown for the four samples (mAb (S): blue, mAb (LR): green, ADC (S): red, ADC (LR): orange). N387 is modified to aspartic acid (Asp) resulting in a mass increase of 1 Da and m/z 1273.06 for the $[M+2H]^{2+}$ state. The extracted chromatograms on the left correspond to the deamidated (red) and native (black) peptide for each sample. The peak of the deamidated peptide can only be seen as a shoulder of the native peptide peak.

Table A5: Deamidation and oxidation sites for mAb (S). The peptides in their observed charge states, the intensities for the MS peaks at full width half maximum (FWHM) and the integrated areas of the chromatogram peaks on the extracted chromatograms are given, with the peak start, end and top. The succinimidyl intermediate is shown where observed. Table 3.8 (main text) shows the relative amounts for each modification.

mAb (S)

Site	Peptide	m/z	Charge state	Comment	RT	Intensity at FWHM	Integrated area	Peak start	Peak end	Peak top
M4	LC1–18	947.96	2	Oxidation	8.14	428	127.85	8.04	8.27	8.11
		939.94	2	No oxidation	10.70	62900	19148.70	10.60	10.83	10.68
M255	HC252–258	851.42	1	Oxidation	5.91	596	167.48	5.80	6.02	5.87
		835.44	1	No oxidation	6.96	33400	8838.33	6.86	7.08	6.94
N55	HC51–59	1085.54	1	Deamidation	6.09	1370	322.90	6.02	6.17	6.09
		1084.54	1	No deamidation	5.91	4810	1490.49	5.82	5.99	5.90
		1067.52	1	Succinimidyl	6.68	750	94.30	6.60	6.79	6.68
N318	HC305–320	904.98	2	Deamidation	17.72	4790	988.64	17.65	17.87	17.73
		904.50	2	No deamidation	17.37	52600	16559.49	17.24	17.53	17.35
		603.67	3	Deamidation	17.72	339	114.02	17.61	17.89	17.79
		603.35	3	No deamidation	17.37	3210	1162.41	17.26	17.53	17.36
		597.67	3	Succinimidyl	17.72	692	211.46	17.61	17.91	17.75
	HC305–323	743.73	3	Deamidation	16.72	19900	4362.58	16.65	16.82	16.72
		743.41	3	No deamidation	16.52	49300	18563.58	16.40	16.63	16.48
		737.73	3	Succinimidyl	16.93	1740	610.02	16.81	17.05	16.91
		1115.10	2	Deamidation	16.72	3090	397.40	16.65	16.80	16.70
		1114.61	2	No deamidation	16.52	8220	2995.68	16.40	16.66	16.50
		1106.09	2	Succinimidyl	16.93	416	104.91	16.80	17.04	16.89
N387	HC374–395	1273.06	2	Deamidation	15.09	412	145.18	15.03	15.10	15.05
		1272.56	2	No deamidation	14.87	6580	1822.10	14.74	14.99	14.85
	tentative	831.69	3	Succinimidyl	15.93	845	192.12	15.79	16.07	15.87

Table A6: Similarly for mAb (LR).

mAb (LR)

Site	Peptide	m/z	Charge state	Comment	RT	Intensity at FWHM	Integrated Area	Peak start	Peak end	Peak top
M4	LC1–18	947.94	2	Oxidation	8.14	314	99.26	8.03	8.24	8.12
		939.94	2	No oxidation	10.70	78100	23027.62	10.60	10.82	10.68
M255	HC252–258	851.42	1	Oxidation	5.91	909	245.98	5.79	6.00	5.87
		835.43	1	No oxidation	6.96	30000	7229.66	6.86	7.08	6.94
N55	HC51–59	1085.53	1	Deamidation	6.09	1310	329.36	6.02	6.17	6.07
		1084.54	1	No deamidation	5.91	4600	1199.33	5.82	5.99	5.91
		1067.52	1	Succinimidyl	6.70	774	93.65	6.59	6.78	6.68
N318	HC305–320	905.00	2	Deamidation	17.74	4150	1385.58	17.62	17.90	17.73
		904.51	2	No deamidation	17.37	54800	18799.74	17.25	17.53	17.35
		603.65	3	Deamidation		361	112.63	17.62	17.84	17.72
		603.35	3	No deamidation		3460	1239.52	17.26	17.53	17.36
		597.67	3	Succinimidyl	17.74	726	263.97	17.62	17.85	17.75
	HC305–323	743.74	3	Deamidation	16.76	16900	4012.60	16.67	16.92	16.73
		743.40	3	No deamidation	16.55	26400	10267.26	16.43	16.66	16.53
		737.73	3	Succinimidyl	16.94	1330	394.87	16.84	17.07	16.92
		1115.09	2	Deamidation		2460	994.07	16.67	16.96	16.72
		1114.61	2	No deamidation		4790	1400.65	16.43	16.67	16.53
		1106.09	2	Succinimidyl	16.94	212	73.13	16.95	17.04	16.87
N387	HC374–395	1273.07	2	Deamidation	15.08	337	272.19	14.97	15.07	15.05
		1272.57	2	No deamidation	14.88	6100	2047.53	14.73	15.05	14.85

Table A7: Similarly for ADC (S).

ADC (S)

Site	Peptide	m/z	Charge state	Comment	RT	Intensity at FWHM	Integrated Area	Peak start	Peak end	Peak top
M4	LC1–18	947.94	2	Oxidation	8.19	102	46.41	7.96	8.33	8.14
		939.95	2	No oxidation	10.72	13000	4487.58	10.65	10.84	10.72
M255	HC252–258	851.46	1	Oxidation	5.94	175	48.84	5.82	6.02	5.91
		835.43	1	No oxidation	6.99	6430	1370.73	6.92	7.09	6.97
N55	HC51–59	1085.56	1	Deamidation	6.12	269	78.85	6.04	6.19	6.11
		1084.55	1	No deamidation	5.94	725	189.03	5.87	6.04	5.92
		1067.54	1	Succinimidyl	6.72	86	25.11	6.61	6.84	6.72
N318	HC305–320	905.00	2	Deamidation	17.72	1360	316.24	17.67	17.89	17.75
		904.51	2	No deamidation	17.42	12000	4709.07	17.27	17.63	17.40
		603.68	3	Deamidation		143	31.22	17.68	17.84	17.74
		603.35	3	No deamidation		882	320.73	17.30	17.56	17.40
		597.67	3	Succinimidyl	17.72	281	62.45	17.62	17.86	17.74
	HC305–323	743.73	3	Deamidation	16.77	4420	702.68	16.70	16.84	16.75
		743.41	3	No deamidation	16.57	9720	3468.06	16.46	16.70	16.55
	tentative	737.73	3	Succinimidyl	16.93	500	153.70	16.86	17.07	16.91
		1115.09	2	Deamidation		678	291.60	16.73	16.84	16.77
		1114.61	2	No deamidation		1240	644.26	16.47	16.75	16.55
N387	HC374–395	1273.08	2	Deamidation	15.09	139	58.96	15.02	15.10	15.05
		1272.56	2	No deamidation	14.89	1520	457.09	14.77	15.02	14.87

Table A8: Similarly for ADC (LR).

ADC (LR)

Site	Peptide	m/z	Charge state	Comment	RT	Intensity at FWHM	Integrated Area	Peak start	Peak end	Peak top
M4	LC1–18	947.94	2	Oxidation	8.16	291	52.41	8.04	8.21	8.16
		939.95	2	No oxidation	10.72	38800	12942.43	10.62	10.83	10.70
M255	HC252–258	851.43	1	Oxidation	5.92	352	96.90	5.81	5.97	5.91
		835.44	1	No oxidation	6.98	17700	4755.08	6.84	7.16	6.96
N55	HC51–59	1085.54	1	Deamidation	6.11	885	206.11	6.02	6.19	6.09
		1084.54	1	No deamidation	5.92	2910	797.92	5.84	6.02	5.91
		1067.52	1	Succinimidyl	6.70	580	68.67	6.60	6.80	6.70
N318	HC305–320	905.01	2	Deamidation	17.74	2510	628.85	17.66	17.92	17.75
		904.51	2	No deamidation	17.40	28900	10091.38	17.25	17.58	17.38
		603.66	3	Deamidation		231	51.70	17.67	17.81	17.79
		603.34	3	No deamidation		1950	656.24	17.28	17.54	17.36
		597.67	3	Succinimidyl	17.74	429	104.23	17.68	17.87	17.73
	HC305–323	743.73	3	Deamidation	16.77	10800	2486.53	16.68	16.85	16.75
		743.41	3	No deamidation	16.55	30100	11372.43	16.43	16.67	16.53
		737.73	3	Succinimidyl	16.94	1590	376.22	16.82	17.05	16.92
		1115.10	2	Deamidation		1220	647.83	16.70	16.82	16.75
		1114.62	2	No deamidation		4300	1948.07	16.43	16.73	16.53
		1106.09	2	Succinimidyl	16.94	214	58.67	16.86	16.99	16.92
N387	HC374–395	1273.06	2	Deamidation	15.11	244	184.32	15.06	15.14	15.11
		1272.56	2	No deamidation	14.89	4220	1311.89	14.75	15.03	14.87

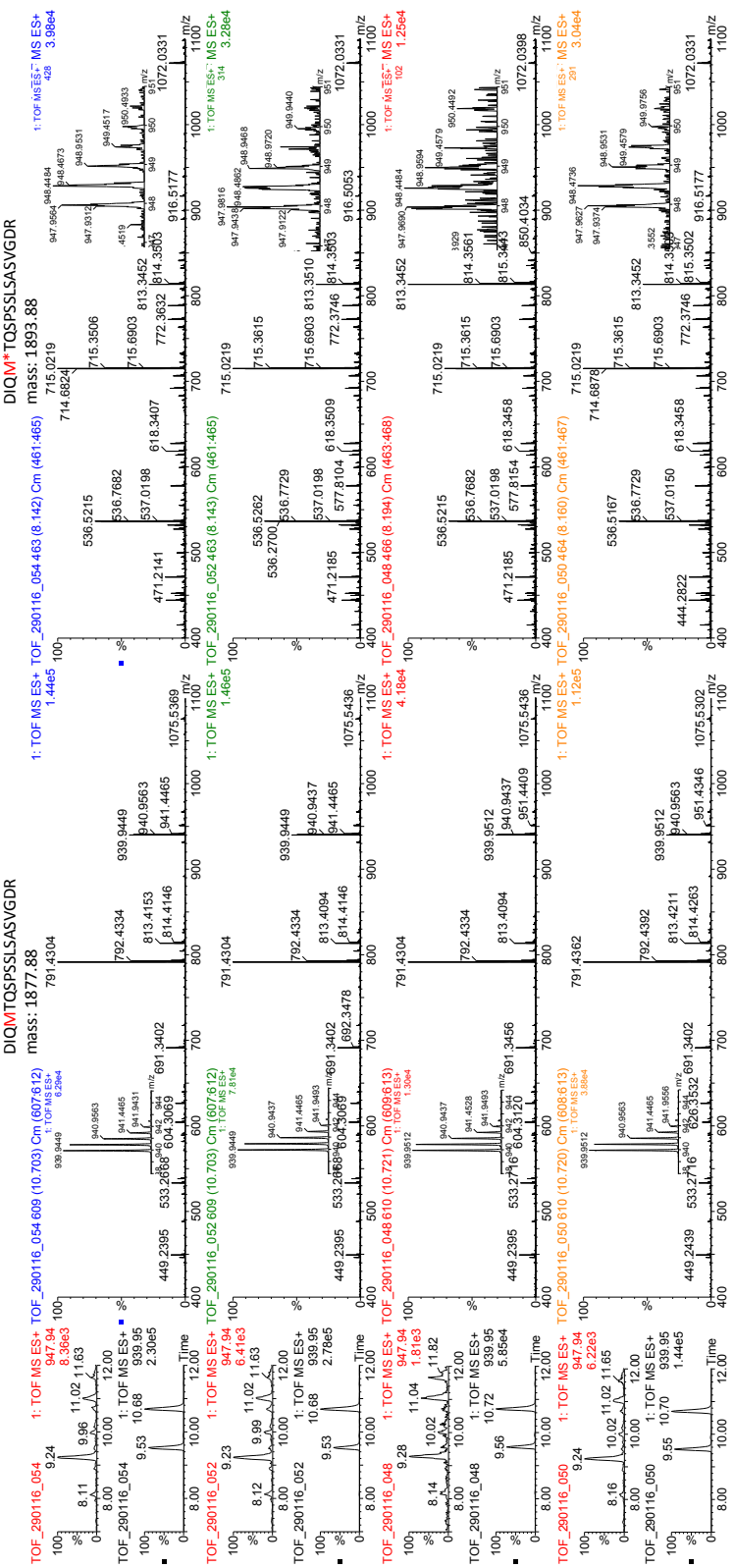


Figure A13: Comparison of the spectra of the native (left) and oxidised (right) peptide LC1–18 with RT ~10.7 min and ~8.1 min respectively are shown for the four samples (mAb (S): blue, mAb (LR): green, ADC (S): red, ADC (LR): orange). M4 is modified to methionine sulfoxide (Mso) resulting in a mass increase of 16 Da and m/z 947.93 for the $[M+2H]^{2+}$ state. The extracted chromatograms on the left correspond to the oxidised (red) and native (black) peptide for each sample.

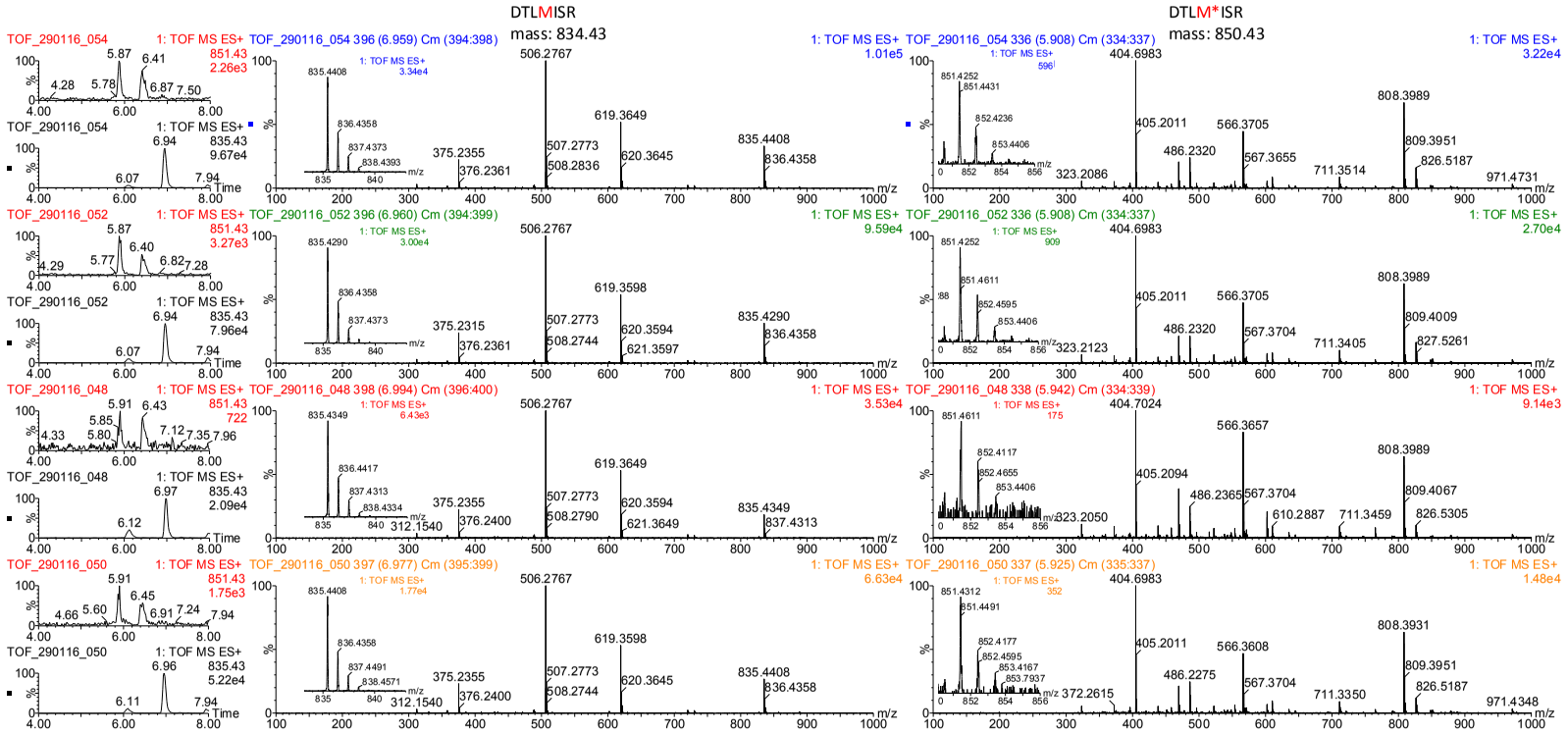


Figure A14: Comparison of the spectra of the native (left) and oxidised (right) peptide HC252–258 with RT ~6.9 min and ~5.9 min respectively are shown for the four samples (mAb (S): blue, mAb (LR): green, ADC (S): red, ADC (LR): orange). M255 is modified to methionine sulfoxide (Mso) resulting in a mass increase of 16 Da and m/z 851.45 for the $[M+H]^+$ state. The extracted chromatograms on the left correspond to the oxidised (red) and native (black) peptide for each sample.

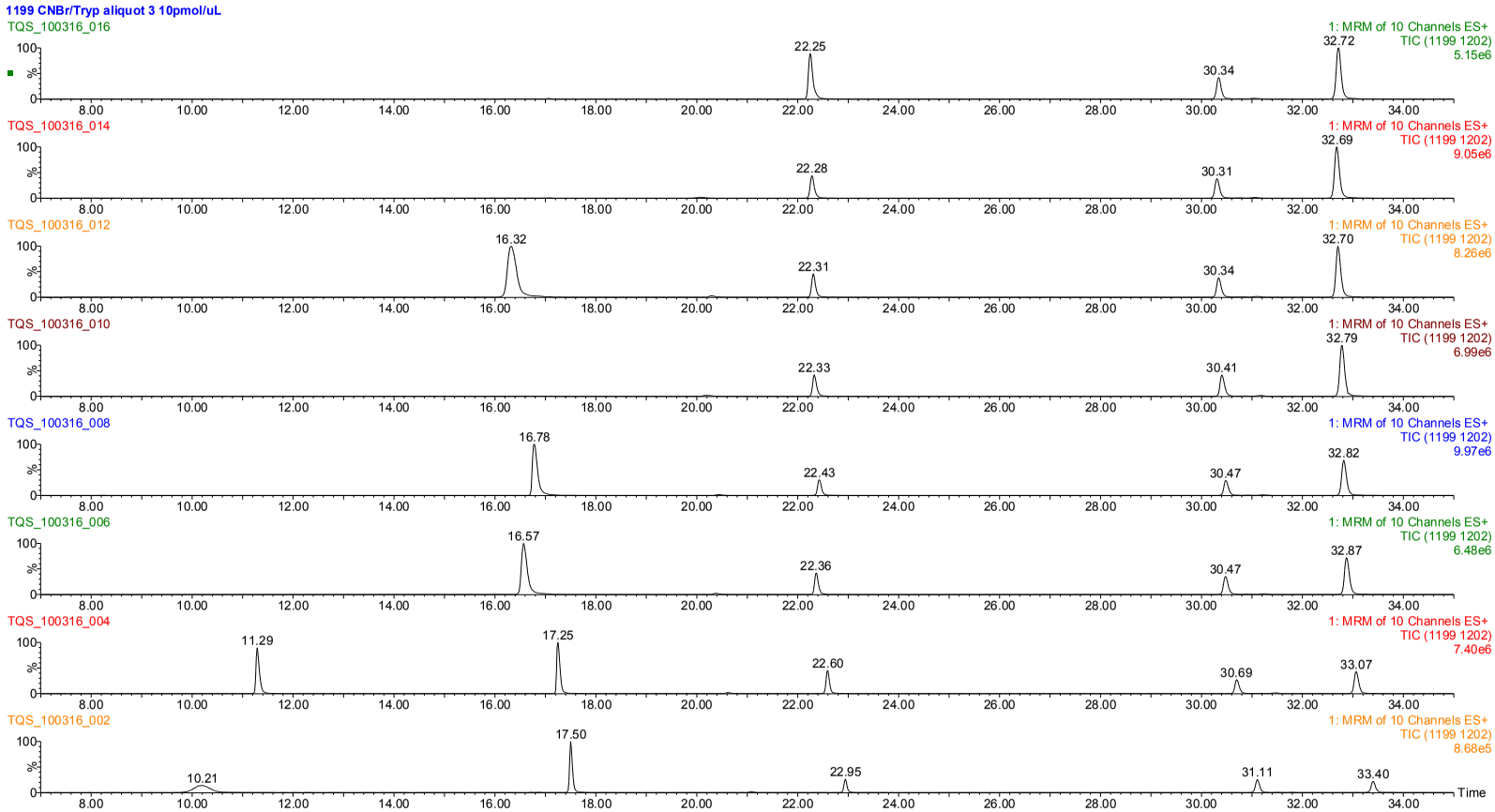


Figure A15: TIC plots from the eight TQ-S runs for the ADC (S). The first and the last runs were excluded, as less consistent. The most abundant and consistent among the repetitions peak (RT ~22.3 min) was used as the marker peptide (LC46–61), while the peaks with RT ~30.3 min and ~32.7 min correspond to the drug-conjugated peptides HC222–225 and LC212–214 respectively. The measured integrated areas are given in Table A9.

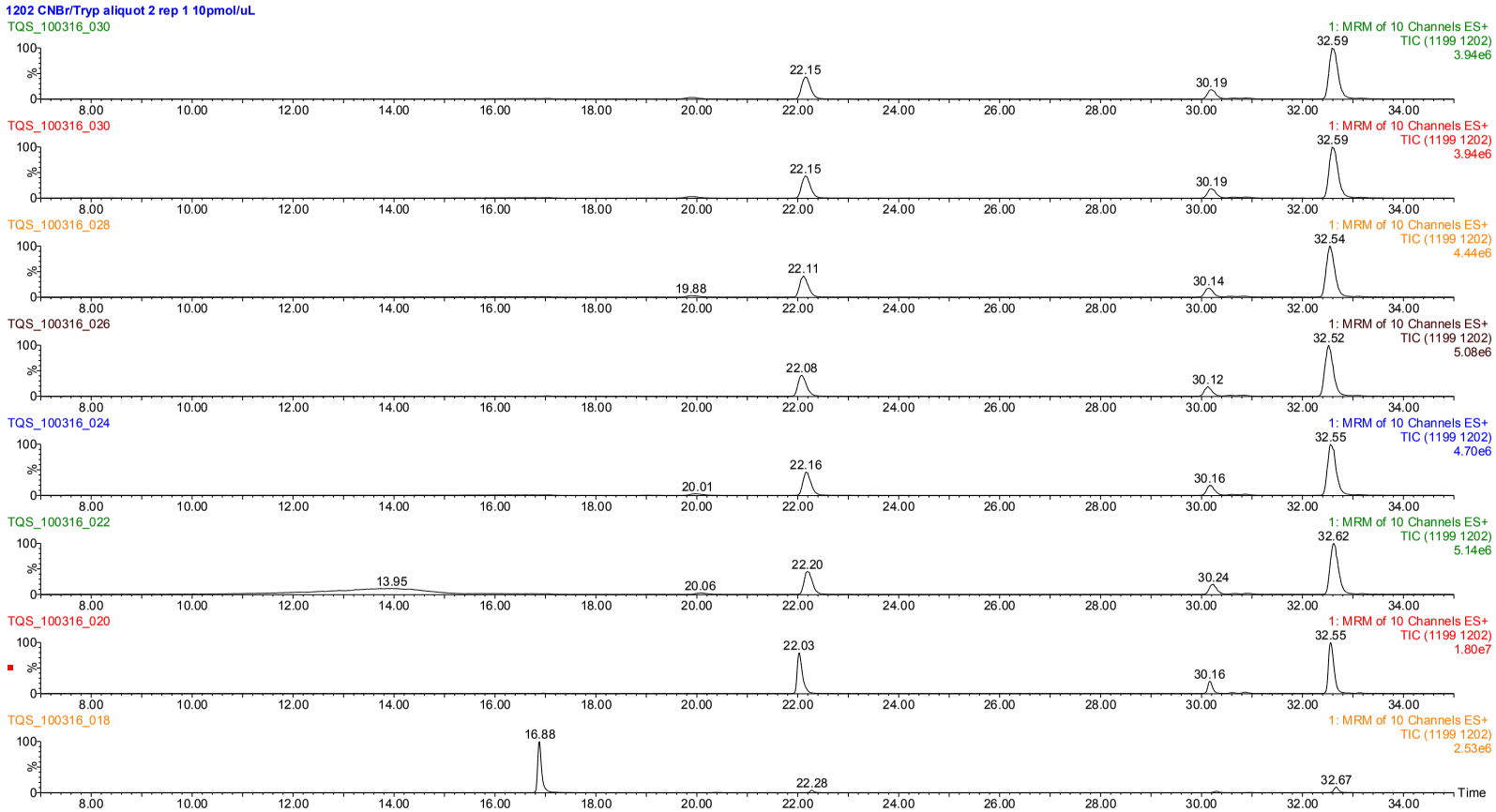


Figure A16: TIC plots from the eight TQ-S runs for the ADC (LR). The first and the last runs were excluded, as less consistent. The most abundant and consistent among the repetitions peak (RT ~22.2 min) was used as the marker peptide (LC46–61), while the peaks with RT ~30.2 min and ~32.6 min correspond to the drug-conjugated peptides HC222–225 and LC212–214 respectively. The measured integrated areas are given in Table A9.

Table A9: Integrated areas for each one of the monitored transitions for the marker and the drug-conjugated peptides for the ADC samples. The peaks from the TIC plot of the TQ-S data were used. The relative abundance of the drug-conjugated peptides, for each sample, was calculated by dividing the sum of the areas from the two transitions by the marker peptide area. The resulting plot is given in Figure 3.15.

ADC (S)

Peptide	Property	Transition	1	2	3	4	5	6
LC46-61	marker	886.99 > 359.20	256912.33	223243.03	271722.06	257371.81	329028.44	361112.22
LC212-214	with drug	812.44 > 686.50	134498.75	206610.17	301530.78	290352.97	343596.25	373547.75
		812.44 > 718.50	191451.77	308971.25	462551.19	460773.38	549256.44	621008.00
HC222-225	with drug	884.48 > 686.50	73861.38	89975.53	106234.76	106264.66	115188.71	122080.96
		884.48 > 718.50	118017.12	151571.92	194995.38	190655.17	211164.73	236677.61

ADC (LR)

Peptide	Property	Transition	1	2	3	4	5	6
LC46-61	marker	886.99 > 359.20	1572686.13	416346.19	369991.47	376137.81	343868.13	328005.53
LC212-214	with drug	812.44 > 686.50	888619.00	300738.09	267279.31	279121.19	246631.67	225018.91
		812.44 > 718.50	1338262.00	628304.25	594221.13	643845.50	579126.56	518342.16
HC222-225	with drug	884.48 > 686.50	166576.53	49744.11	45056.84	44440.23	39547.89	36154.06
		884.48 > 718.50	289689.78	123203.20	116556.29	118799.48	105760.67	103411.59

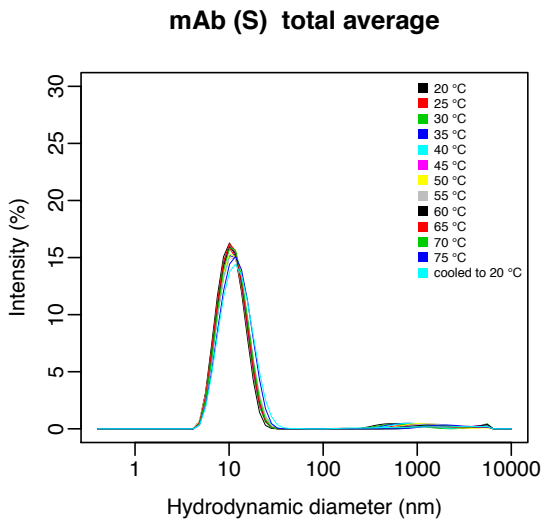


Figure A17: Size distribution (presented in intensity values, nm) for mAb (S) at a temperature range of 20–75°C and after the sample was cooled down to 20°C. The distribution becomes wider and slightly shifts to larger values at the end of the experiment. The presented data are averaged over 6 replicates.

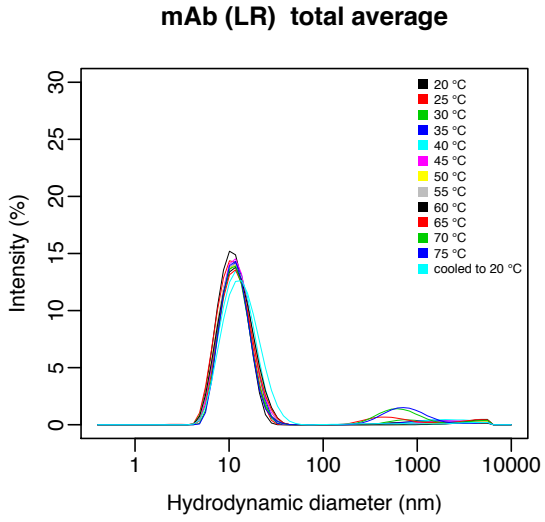


Figure A18: Size distribution (presented in intensity values, nm) for mAb (LR) at a temperature range of 20–75°C and after the sample was cooled down to 20°C. The distribution becomes wider and slightly shifts to larger values at the end of the experiment. The presented data are averaged over 6 replicates.

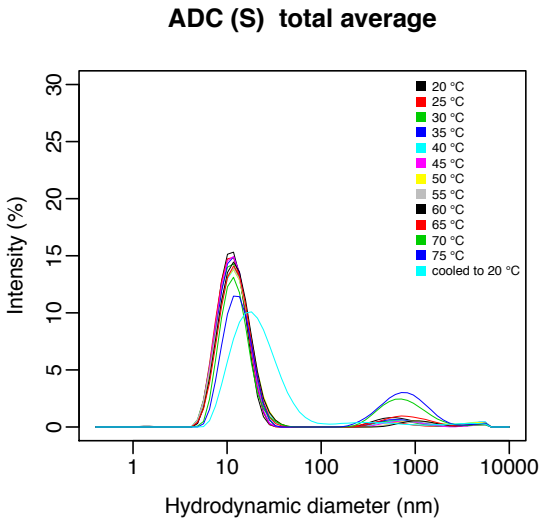


Figure A19: Size distribution (presented in intensity values, nm) for ADC (S) at a temperature range of 20–75°C and after the sample was cooled down to 20°C. The increased width, the shift to larger values and the loss in intensity of the distribution at the end of the experiment, indicate the appearance of larger particles. The presented data are averaged over 6 replicates.

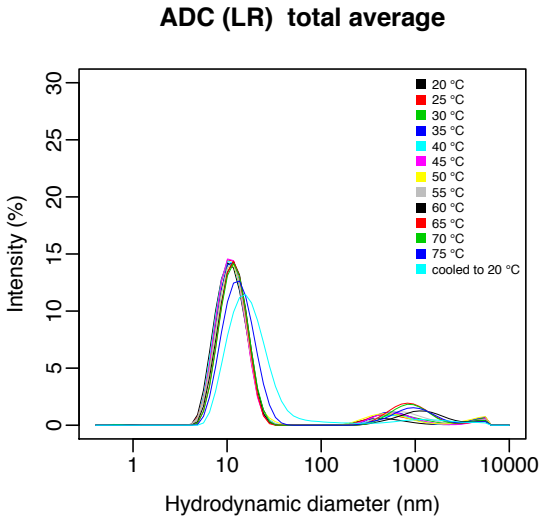


Figure A20: Size distribution (presented in intensity values, nm) for ADC (LR) at a temperature range of 20–75°C and after the sample was cooled down to 20°C. The increased width, the shift to larger values and the loss in intensity of the distribution at the end of the experiment, indicate the appearance of larger particles. The presented data are averaged over 6 replicates.

Table A10: Overall mean particle sizes presented in terms of number values (nm) and the standard deviations (over six replicates) with a 95% confidence interval are shown for each sample at a temperature range of 20–75°C and after the samples were cooled down to 20°C.

Sample	20°C	25°C	30°C	35°C	40°C	45°C	50°C	55°C	60°C	65°C	70°C	75°C	Cooled to 20°C
mAb (S)	6.8±0.5	6.8±0.9	6.9±0.6	7.0±1.1	6.9±1.5	6.9±0.6	6.9±1.5	7.1±2.3	7.1±1.3	6.9±0.7	7.1±1.3	7.3±1.6	7.1±0.7
mAb (LR)	6.7±0.6	6.5±0.5	7.0±2.3	7.3±2.8	7.4±3.4	7.5±2.7	7.0±2.1	7.1±1.9	7.0±1.7	7.3±1.2	7.4±0.7	7.5±0.7	7.1±0.9
ADC (S)	7.2±1.1	6.9±0.9	6.9±1.2	7.3±1.4	7.0±1.6	7.1±1.3	7.2±1.1	7.0±1.4	7.5±2.0	7.5±2.1	8.1±1.4	8.3±1.0	8.9±0.6
ADC (LR)	6.6±0.9	7.0±1.7	7.0±1.8	7.1±1.8	6.8±1.0	7.1±1.9	7.3±2.3	7.4±2.4	7.8±2.6	7.9±2.3	7.9±2.1	7.9±1.7	8.5±1.1

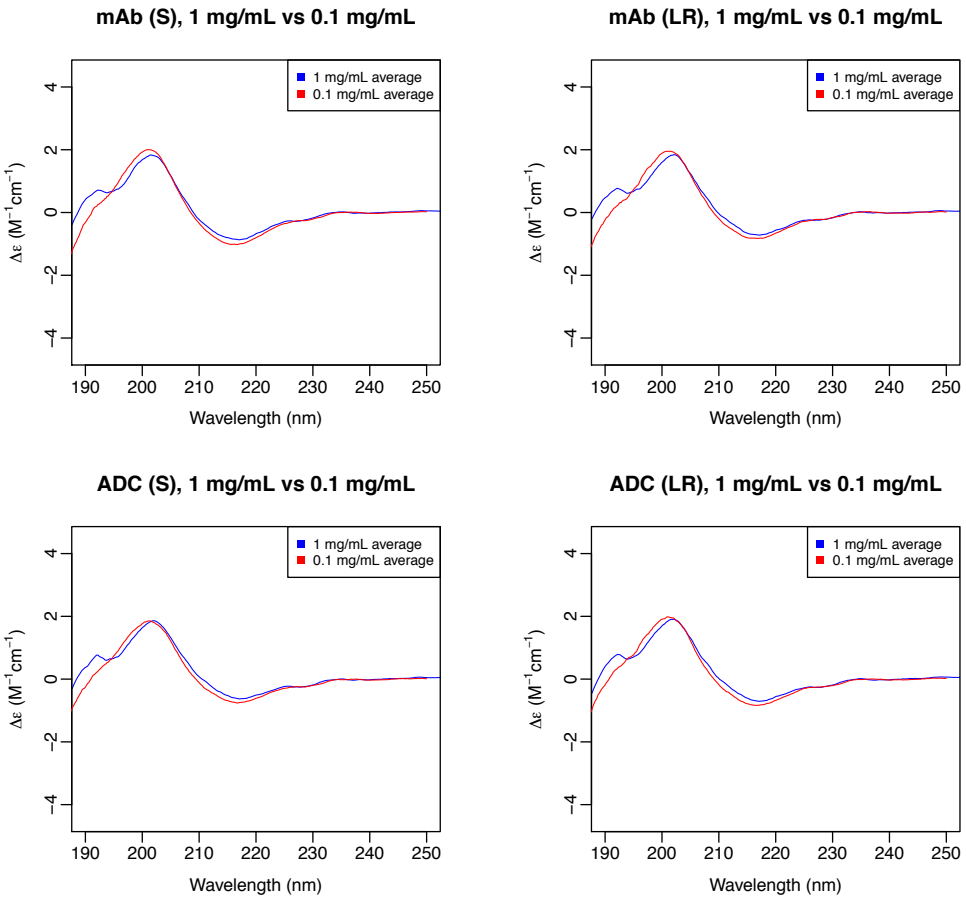


Figure A21: Comparison of the CD spectra of the four trastuzumab samples at their original concentration (1 mg/mL) using a 0.01 cm demountable cuvette (blue) and after 10-fold dilution (0.1 mg/mL) using a 0.1 cm fixed cuvette (red). All data were acquired at 20 °C. The high concentration spectra are the average of 5 replicates (16 scans each), and the diluted spectra are averaged over 6 replicates (16 scans each). The 1 mg/mL spectra were recorded on a J-815 spectrophotometer, while J-1500 was used for the data at 0.1 mg/mL. In order to correct for the variance in the calibration between the two instruments, the high concentration spectra were shifted by −3 nm.

Table A11: Maximum, zero-crossing and minimum points/values of the normalised CD spectra for the 6 replicates. The samples were 10-fold diluted and a 0.1 cm fixed path length cuvette was used. The spectra were recorded at 20 °C and each spectrum is the accumulation of 16 scans.

Sample	Rep	Max (nm)	Max value (Δε)	Zero-crossing (nm)	Min (nm)	Min value (Δε)
mAb (S)	1	201.0	2.00	208.4	215.6	−1.12
	2	200.8	1.97	208.4	216.8	−1.04
	3	201.0	2.11	208.8	216.8	−1.02
	4	200.4	2.02	208.2	214.4	−0.96
	5	201.6	2.02	208.2	216.4	−1.04
	6	202.0	1.98	208.4	216.8	−1.05
mAb (LR)	1	200.8	1.97	209.0	216.8	−0.86
	2	201.8	2.01	209.2	216.4	−0.78
	3	201.6	2.04	209.4	216.0	−0.80
	4	201.2	1.94	208.8	216.6	−0.85
	5	200.6	1.96	208.6	217.8	−0.87
	6	201.2	1.99	208.8	216.8	−0.87
ADC (S)	1	201.0	1.94	209.2	218.2	−0.79
	2	201.4	1.93	209.8	217.6	−0.70
	3	201.2	1.85	209.6	216.6	−0.76
	4	200.4	1.85	209.2	216.8	−0.77
	5	201.0	1.84	209.4	216.6	−0.77
	6	201.4	1.83	208.8	215.8	−0.86
ADC (LR)	1	201.0	1.96	209.0	217.2	−0.90
	2	201.0	1.94	209.2	217.0	−0.89
	3	201.0	2.04	209.0	216.0	−0.83
	4	201.0	2.08	209.0	215.8	−0.81
	5	201.4	1.97	209.0	217.0	−0.77
	6	202.0	1.99	208.8	216.4	−0.91

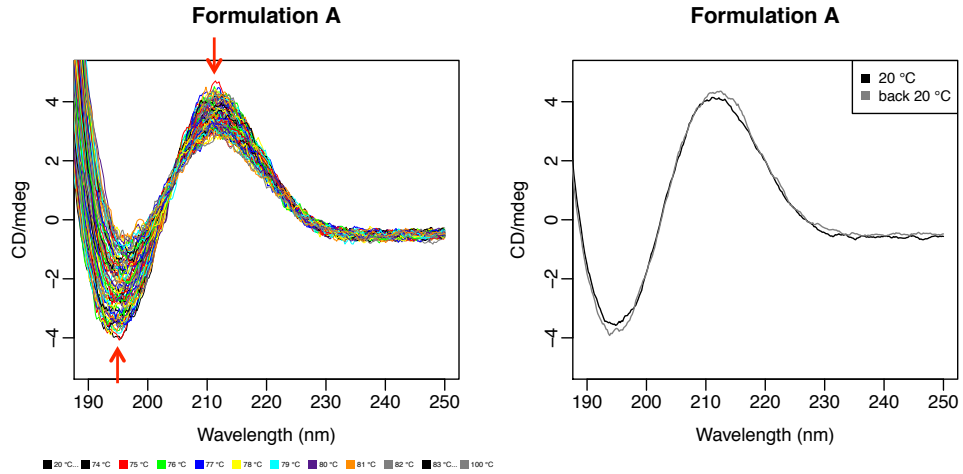


Figure A22: Left: CD spectra of the formulation after 10-fold dilution at a temperature range of 20–100 °C, per 1 °C. The spectra were recorded on a J-1500 spectrophotometer, using a 0.1 cm fixed path length cuvette according to the temperature experiment for the samples. The red arrows show the direction of the decreasing signal intensities at 195 nm and 212 nm with increasing temperature. Right: Comparison of the formulation spectra at 20 °C and at the end of the temperature experiment, after the cuvette was cooled down to 20 °C.

Table A12: Unfolding temperatures for the trastuzumab samples extracted from the CD spectra of each replicate. The averaged temperatures for the first and the last three replicates are given as well as the total averages for each sample, also described in the text.

Sample	Replicate	Unfolding Temperature	Average	
mAb (S)	1	82 °C		
	2	80 °C	rep 1–3	82 °C
	3	82 °C		
	4	82 °C		
	5	82 °C	rep 4–6	82 °C
	6	82 °C		
			Total	82 °C
mAb (LR)	1	83 °C		
	2	81 °C	rep 1–3	82 °C
	3	82 °C		
	4	82 °C		
	5	83 °C	rep 4–6	83 °C
	6	82 °C		
			Total	82 °C
ADC (S)	1	80 °C		
	2	78 °C	rep 1–3	79 °C
	3	78 °C		
	4	80 °C		
	5	80 °C	rep 4–6	81 °C
	6	81 °C		
			Total	80 °C
ADC (LR)	1	82 °C		
	2	80 °C	rep 1–3	80 °C
	3	79 °C		
	4	78 °C		
	5	82 °C	rep 4–6	81 °C
	6	82 °C		
			Total	80 °C

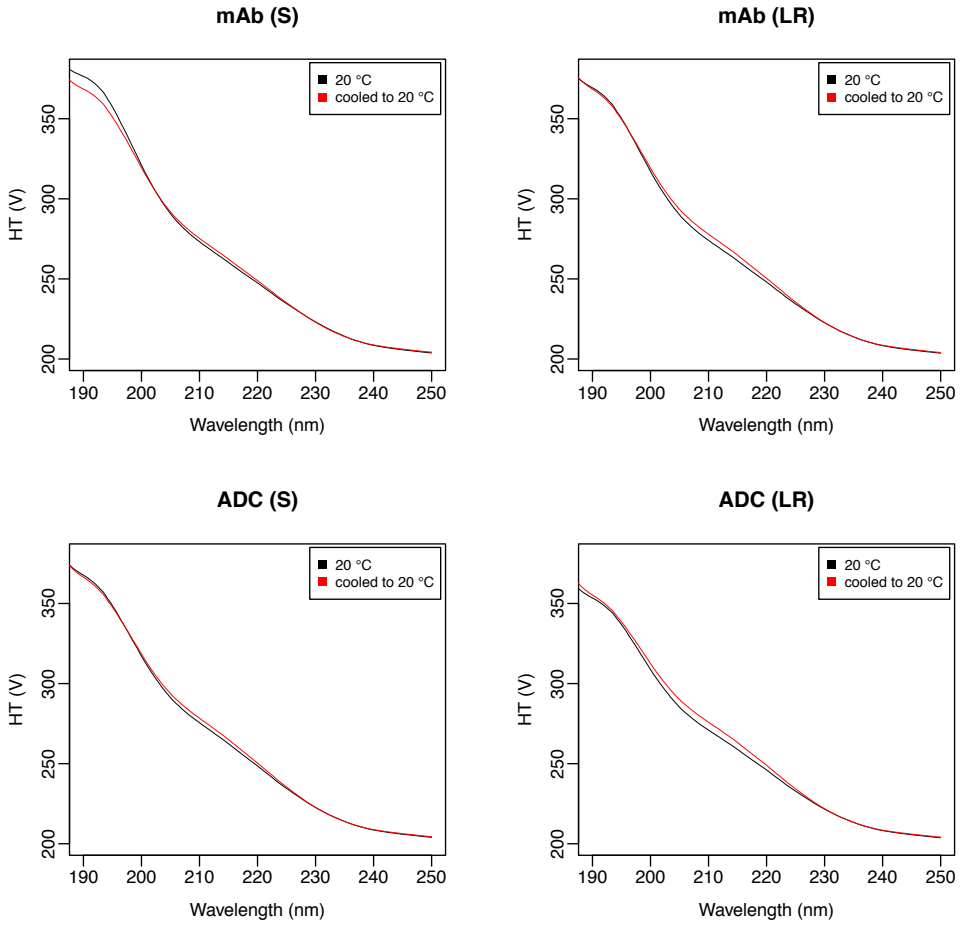


Figure A23: Comparison of the HT spectra of the four trastuzumab samples at 20 °C and at the end of the experiment, after the samples were cooled down to 20 °C. Only the last 3 replicates were used for calculating the average.

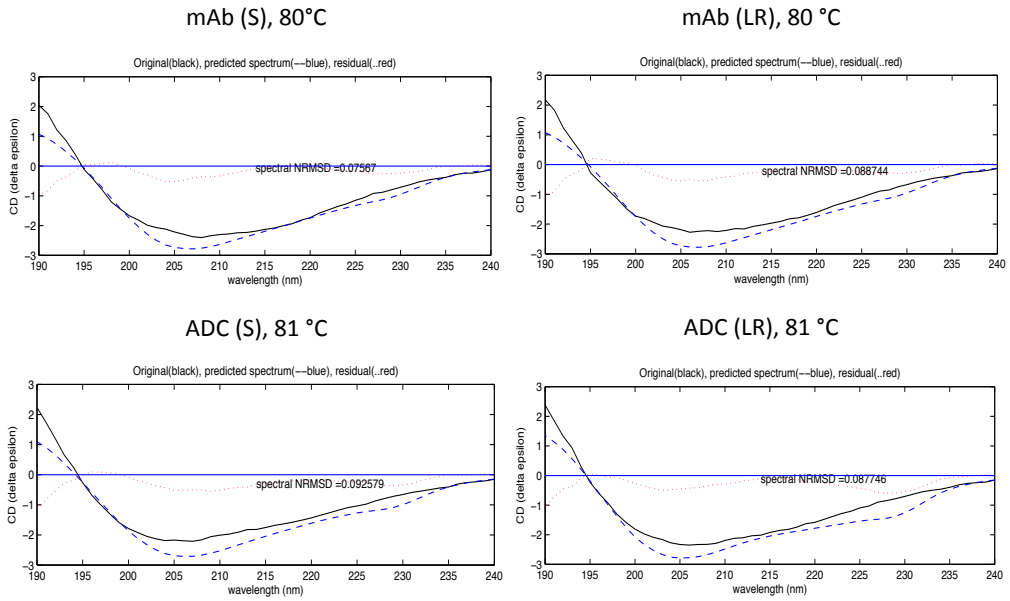


Figure A24: SSNN output from the analysis of the averaged trastuzumab spectra at a temperature range 20–100 °C, per 1 °C. The spectral fits with the NRMSD values at 80 °C and 81 °C are shown for the native and the conjugated antibodies respectively. The incorrectly estimated feature at 228 nm accounts for the misleading increase in α-helix.

Appendix B

PROTEIN H

Table B1: Target wavelengths for the HT maxima for each one of the calibration tests, measured wavelengths at the beginning of the experiment and the respective deviations are presented. The measured values should be ±0.8 nm from the target values.

Test	Target Wavelength (nm)	Wavelength (nm)	Deviation (nm)
Neodymium filter	586.0	585.4	−0.6
Holmium filter	637.5	637.6	0.1
	536.6	536.0	−0.6
	460.2	459.8	−0.4
	453.7	453.6	−0.1
	445.7	445.4	−0.3
	360.9	360.8	−0.1
	333.9	333.8	−0.1
	287.6	287.4	−0.2
	279.4	279.2	−0.2
Benzene vapour	259.0	258.9	−0.1
	253.0	252.8	−0.2
	247.2	247.0	−0.2
	241.7	241.4	−0.3
Ammonia vapour	208.6	208.1	−0.5
	204.8	204.2	−0.6
	201.1	200.4	−0.7
	197.5	196.8	−0.7
	194.0	193.4	−0.6
	190.6	190.1	−0.5
	187.2	186.7	−0.5

Table B2: The maximum target intensity values (mdeg) at the specific wavelengths and the measured intensities before the experiment are shown for the performed tests. The measured values should be ±1 mdeg from the target values. Values that did not meet the criterion are noted in red.

Test	Wavelength (nm)	Target Intensity (mdeg)	Intensity (mdeg)
ACS solution (1 cm cuvette)	291	188.0	189.5
ACS solution (1 mm cuvette)	291	18.8	19.2
	193	−39.1	−39.9
(*R*,*R*)-Na[Co^{III}(EDDS)]	211.5	−50.0	−52.3
(*S*,*S*)-Na[Co^{III}(EDDS)]	211.5	50.0	51.9

SOMATROPIN (RECOMBINANT HUMAN GROWTH HORMONE)

Table B3: Target wavelengths for the HT maxima for each one of the calibration tests, measured wavelengths at the beginning of the experiment and the respective deviations are presented. The measured values should be ±0.8 nm from the target values. Values that did not meet the criterion are marked in red.

Test	Target Wavelength (nm)	Wavelength (nm)	Deviation (nm)
Neodymium filter	586.0	585.1	−0.9
Holmium filter	637.5	637.0	−0.5
	536.6	535.6	−1.0
	460.2	459.4	−0.8
	453.7	453.2	−0.5
	445.7	445.2	−0.5
	360.9	360.6	−0.3
	333.9	333.8	−0.1
	287.6	287.4	−0.2
	279.4	279.2	−0.2
Benzene vapour	259.0	258.8	−0.2
	253.0	252.8	−0.2
	247.2	247.0	−0.2
	241.7	241.4	−0.3
Ammonia vapour	208.6	208.0	−0.6
	204.8	204.2	−0.6
	201.1	200.4	−0.7
	197.5	196.8	−0.7
	194.0	193.4	−0.6
	190.6	190.1	−0.5
	187.2	186.6	−0.6

Table B4: The maximum target intensity values (mdeg) at the specific wavelengths and the measured intensities before the experiment are shown for the performed tests. The measured values should be ±1 mdeg from the target values. Values that did not meet the criterion are noted in red.

Test	Wavelength (nm)	Target Intensity (mdeg)	Intensity (mdeg)
ACS solution (1 cm cuvette)	291	188.0	188.1
ACS solution (1 mm cuvette)	291	18.8	19.2
	193	−39.1	−40.5
(*R,R*)-Na[Co^{III}(EDDS)]	211.5	−50.0	−49.5
(*S,S*)-Na[Co^{III}(EDDS)]	211.5	50.0	51.4

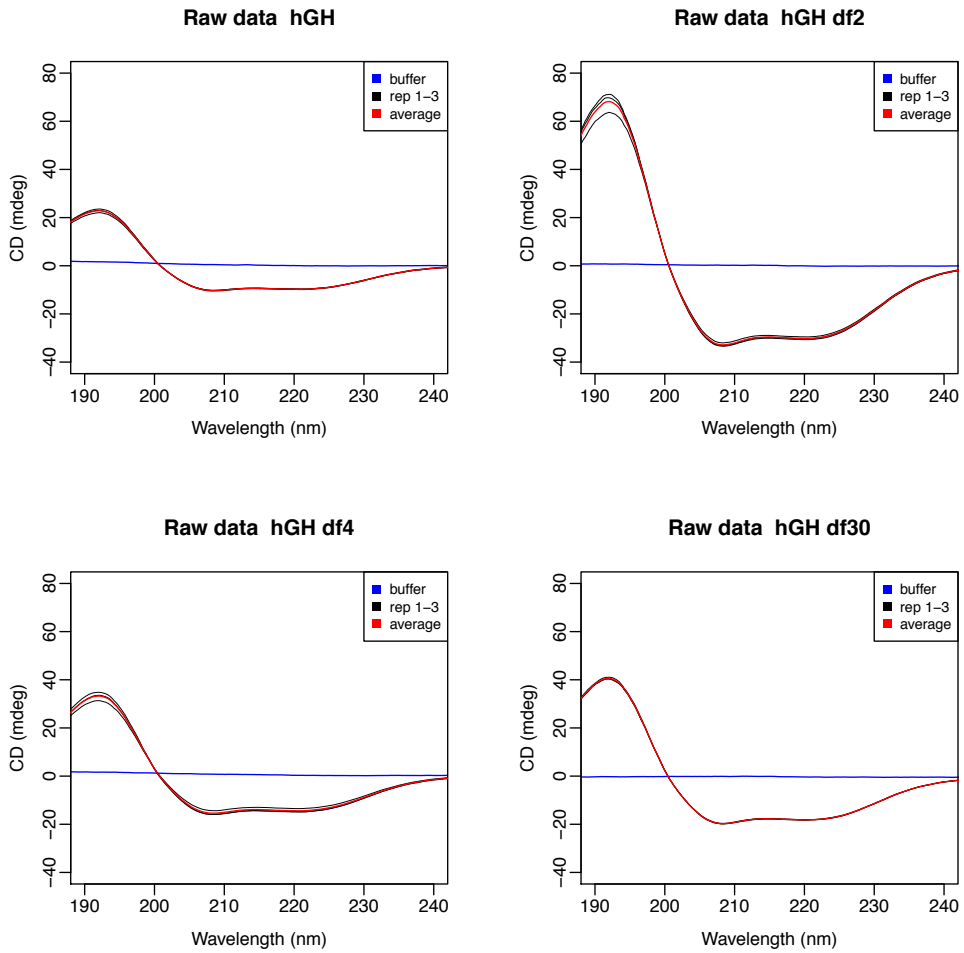


Figure B1: Raw CD data (mdeg) of hGH and reference buffer in the original concentration (nominally 3.56 mg/mL, top left), after 2-fold dilution (top right), 4-fold dilution (bottom left) and 30-fold dilution (bottom right). The three replicates (in black), the averaged spectra (in red) of the sample and the respective buffer spectrum (in blue) are shown.

Table B5: Maximum, zero-crossing and minimum points/values of the normalised CD spectra (Δε units) for the 3 replicates of hGH in different dilutions and after the path length correction was applied.

Sample	Rep	Max (nm)	Max value (Δε)	Zero-crossing (nm)	Min (nm)	Min value (Δε)	Min (nm)	Min value (Δε)
hGH	1	192.2	12.59	200.4	208.2	−6.33	221.0	−5.75
	2	192.2	12.19	200.4	208.2	−6.15	220.2	−5.54
	3	192.2	13.08	200.6	208.2	−6.31	218.6	−5.72
hGH df2	1	192.0	12.09	200.4	208.4	−6.22	219.4	−5.69
	2	192.2	13.57	200.4	208.6	−6.43	219.8	−5.86
	3	192.0	13.30	200.4	208.4	−6.38	219.4	−5.80
hGH df4	1	192.0	11.56	200.4	208.4	−5.78	220.2	−5.27
	2	192.2	12.51	200.4	208.4	−6.30	218.2	−5.73
	3	192.0	12.94	200.4	208.2	−6.36	219.2	−5.77
hGH df30	1	192.0	13.31	200.4	208.2	−6.29	220.0	−5.72
	2	191.8	13.12	200.4	208.2	−6.23	220.0	−5.63
	3	192.0	13.13	200.4	208.4	−6.21	219.2	−5.63

Sample	Program	Ref. set	Helix	Sheet	Turns	Other	NRMSD
hGH	CDSSTR	1	0.61	0.09	0.12	0.19	0.027
	SELCON3		0.57	0.07	0.16	0.21	0.151
	CDSSTR	2	0.62	0.03	0.11	0.24	0.025
	SELCON3		0.62	0.06	0.11	0.21	0.078
	CDSSTR	3	0.63	0.09	0.09	0.19	0.025
	SELCON3		0.58	0.07	0.14	0.22	0.147
	CDSSTR	4	0.61	0.07	0.09	0.23	0.034
	SELCON3		0.56	0.06	0.13	0.25	0.169
	CDSSTR	5	0.61	0.06	0.11	0.22	0.036
	SELCON3		0.58	0.07	0.09	0.27	0.153
	CDSSTR	6	0.65	0.10	0.12	0.14	0.026
	SELCON3		0.59	0.18	0.22	0.01	0.142
	CDSSTR	7	0.61	0.07	0.09	0.23	0.034
	SELCON3		0.59	0.16	0.25	0.01	0.148
	SSNN		0.54	0.07	0.15	0.24	0.020
hGH df2	CDSSTR	1	0.62	0.09	0.12	0.17	0.036
	SELCON3		0.60	0.15	0.22	0.03	0.149
	CDSSTR	2	0.65	0.03	0.11	0.21	0.019
	SELCON3		0.64	0.13	0.15	0.08	0.092
	CDSSTR	3	0.66	0.07	0.10	0.18	0.029
	SELCON3		0.61	0.17	0.22	0.00	0.143
	CDSSTR	4	0.61	0.08	0.09	0.22	0.040
	SELCON3		0.59	0.16	0.25	0.00	0.151
	CDSSTR	5	0.64	0.05	0.10	0.22	0.029
	SELCON3		0.59	0.07	0.09	0.26	0.139
	CDSSTR	6	0.65	0.08	0.10	0.17	0.025
	SELCON3		0.60	0.17	0.22	0.00	0.150
	CDSSTR	7	0.62	0.06	0.08	0.24	0.032
	SELCON3		0.60	0.16	0.25	0.00	0.157
	SSNN		0.57	0.06	0.14	0.23	0.021
hGH df4	CDSSTR	1	0.61	0.09	0.13	0.18	0.037
	SELCON3		0.57	0.17	0.23	0.03	0.147
	CDSSTR	2	0.61	0.03	0.12	0.24	0.023
	SELCON3		0.62	0.14	0.16	0.08	0.089
	CDSSTR	3	0.65	0.07	0.11	0.18	0.030
	SELCON3		0.58	0.18	0.23	0.01	0.130
	CDSSTR	4	0.59	0.08	0.10	0.23	0.044
	SELCON3		0.57	0.16	0.25	0.02	0.147
	CDSSTR	5	0.65	0.03	0.10	0.22	0.033
	SELCON3		0.57	0.07	0.09	0.27	0.144
	CDSSTR	6	0.63	0.07	0.14	0.16	0.031
	SELCON3		0.58	0.18	0.22	0.02	0.141
	CDSSTR	7	0.60	0.07	0.09	0.25	0.035
	SELCON3		0.57	0.16	0.25	0.02	0.155
	SSNN		0.53	0.07	0.16	0.23	0.020
hGH df30	CDSSTR	1	0.62	0.09	0.13	0.17	0.032
	SELCON3		0.59	0.16	0.22	0.03	0.145
	CDSSTR	2	0.62	0.04	0.10	0.25	0.023
	SELCON3		0.64	0.13	0.15	0.09	0.096
	CDSSTR	3	0.63	0.09	0.09	0.19	0.019
	SELCON3		0.60	0.17	0.22	0.00	0.151
	CDSSTR	4	0.61	0.08	0.09	0.22	0.043
	SELCON3		0.58	0.17	0.25	0.01	0.151
	CDSSTR	5	0.63	0.06	0.10	0.21	0.045
	SELCON3		0.59	0.07	0.09	0.26	0.146
	CDSSTR	6	0.65	0.08	0.11	0.16	0.023
	SELCON3		0.60	0.17	0.22	0.00	0.159
	CDSSTR	7	0.61	0.07	0.09	0.23	0.035
	SELCON3		0.59	0.16	0.25	0.00	0.156
	SSNN		0.57	0.06	0.14	0.23	0.023

Table B6: Secondary structure estimates and NRMSD values from analysis of the CD spectra of hGH in different concentrations. The results from CDSSTR, SELCON3 (using reference sets 1–7) and SSNN (reference set 7) are compared. All programs and reference sets presented very close values for hGH in different concentrations, resulting in an overall accurate estimation. The averaged values and standard deviations from the 15 analyses were calculated for each sample and are presented in Table 4.8 in the main text.

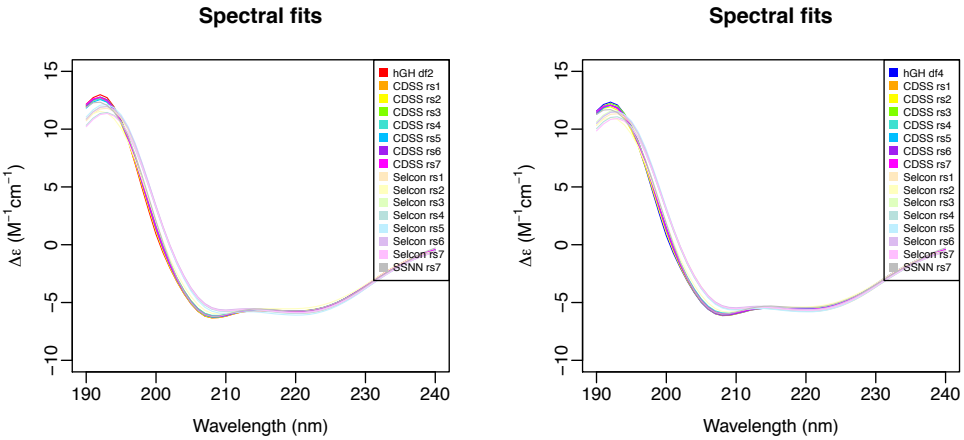


Figure B2: Original and calculated CD spectra from the analysis of hGH after 2- (left) and 4-fold dilution (right) with different software and reference sets. Spectral fits of the calculated spectra are in agreement with each other and with the respective original spectra, suggesting an overall accurate secondary structure analysis.

PROTEIN A

Table B7: Target wavelengths for the HT maxima for each one of the calibration tests, measured wavelengths at the beginning of the experiment and the respective deviations are presented. The measured values should be ±0.8 nm from the target values.

Test	Target Wavelength (nm)	Wavelength (nm)	Deviation (nm)
Neodymium filter	586.0	585.5	−0.5
Holmium filter	637.5	638.0	0.5
	536.6	536.2	−0.4
	460.2	459.8	−0.4
	453.7	453.6	−0.1
	445.7	445.4	−0.3
	360.9	360.8	−0.1
	333.9	333.8	−0.1
	287.6	287.4	−0.2
	279.4	279.0	−0.4
Benzene vapour	259.0	258.8	−0.2
	253.0	252.8	−0.2
	247.2	246.9	−0.3
	241.7	241.4	−0.3
Ammonia vapour	208.6	208.0	−0.6
	204.8	204.1	−0.7
	201.1	200.3	−0.8
	197.5	196.8	−0.7
	194.0	193.3	−0.7
	190.6	190.0	−0.6
	187.2	186.6	−0.6

Table B8: The maximum target intensity values (mdeg) at the specific wavelengths and the measured intensities before the experiment are shown for the performed tests. The measured values should be ±1 mdeg from the target values. Values that did not meet the criterion are noted in red.

Test	Wavelength (nm)	Target Intensity (mdeg)	Intensity (mdeg)
ACS solution (1 cm cuvette)	291	188.0	179.6
ACS solution (1 mm cuvette)	291	18.8	17.9
	193	−39.1	−37.5
(*R,R*)-Na[Co^{III}(EDDS)]	211.5	−50.0	−48.0
(*S,S*)-Na[Co^{III}(EDDS)]	211.5	50.0	49.4

ETANERCEPT

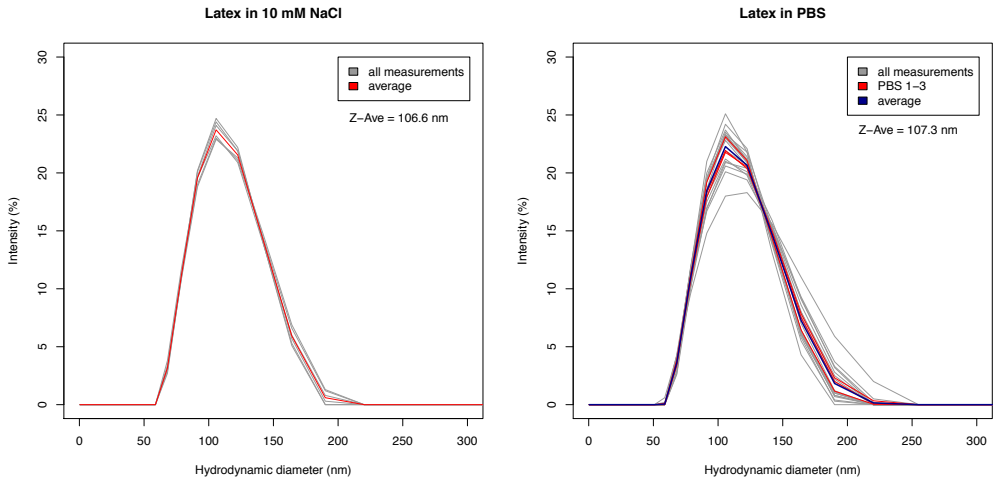


Figure B3: Left: Particle distribution by intensity of 100 nm latex spheres measured in 10 mM NaCl solution. Right: Particle distribution by intensity of 100 nm latex spheres measured in PBS buffer solution. Individual measurements are shown in grey, replicates (averaged over 6 measurements) are shown in red and overall average over three replicates is shown in blue.

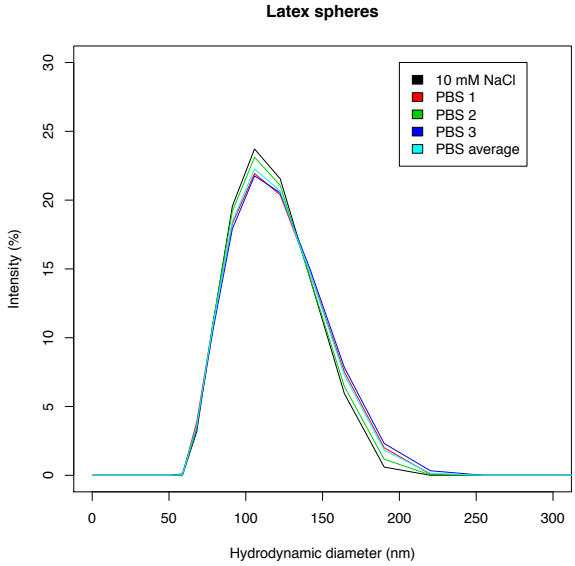


Figure B4: Average particle distributions by intensity of 100 nm latex spheres in NaCl (black) and PBS replicates 1–3 (red, green and blue) solutions averaged over 6 measurements. The overall average distribution for PBS over 3 replicates (18 measurements in total) is shown in cyan.

Table B9: Target wavelengths for the HT maxima for each one of the calibration tests, measured wavelengths at the beginning of the experiment and the respective deviations are presented. The measured values should be ±0.8 nm from the target values. Values that did not meet the criterion are marked in red.

Test	Target Wavelength (nm)	Wavelength (nm)	Deviation (nm)
Neodymium filter	586.0	584.4	−1.6
Holmium filter	637.5	637.8	0.3
	536.6	535.4	−1.2
	460.2	459.3	−0.9
	453.7	453.0	−0.7
	445.7	444.8	−0.9
	360.9	360.1	−0.8
	333.9	333.1	−0.8
	287.6	287.0	−0.6
	279.4	278.9	−0.5
Benzene vapour	259.0	258.6	−0.4
	253.0	252.5	−0.5
	247.2	246.7	−0.5
	241.7	241.2	−0.5
Ammonia vapour	208.6	207.6	−1.0
	204.8	203.7	−1.1
	201.1	200.0	−1.1
	197.5	196.4	−1.1
	194.0	193.0	−1.0
	190.6	189.6	−1.0
	187.2	186.4	−0.8

Table B10: The maximum target intensity values (mdeg) at the specific wavelengths and the measured intensities before the experiment are shown for the performed tests. The measured values should be ±1 mdeg from the target values. Values that did not meet the criterion are noted in red.

Test	Wavelength (nm)	Target Intensity (mdeg)	Intensity (mdeg)
ACS solution (1 cm cuvette)	291	188.0	192.5
ACS solution (1 mm cuvette)	291	18.8	19.1
	193	−39.1	−39.3
(*R,R*)-Na[Co^{III}(EDDS)]	211.5	−50.0	−53.3
(*S,S*)-Na[Co^{III}(EDDS)]	211.5	50.0	54.1

DLS Spectroscopy

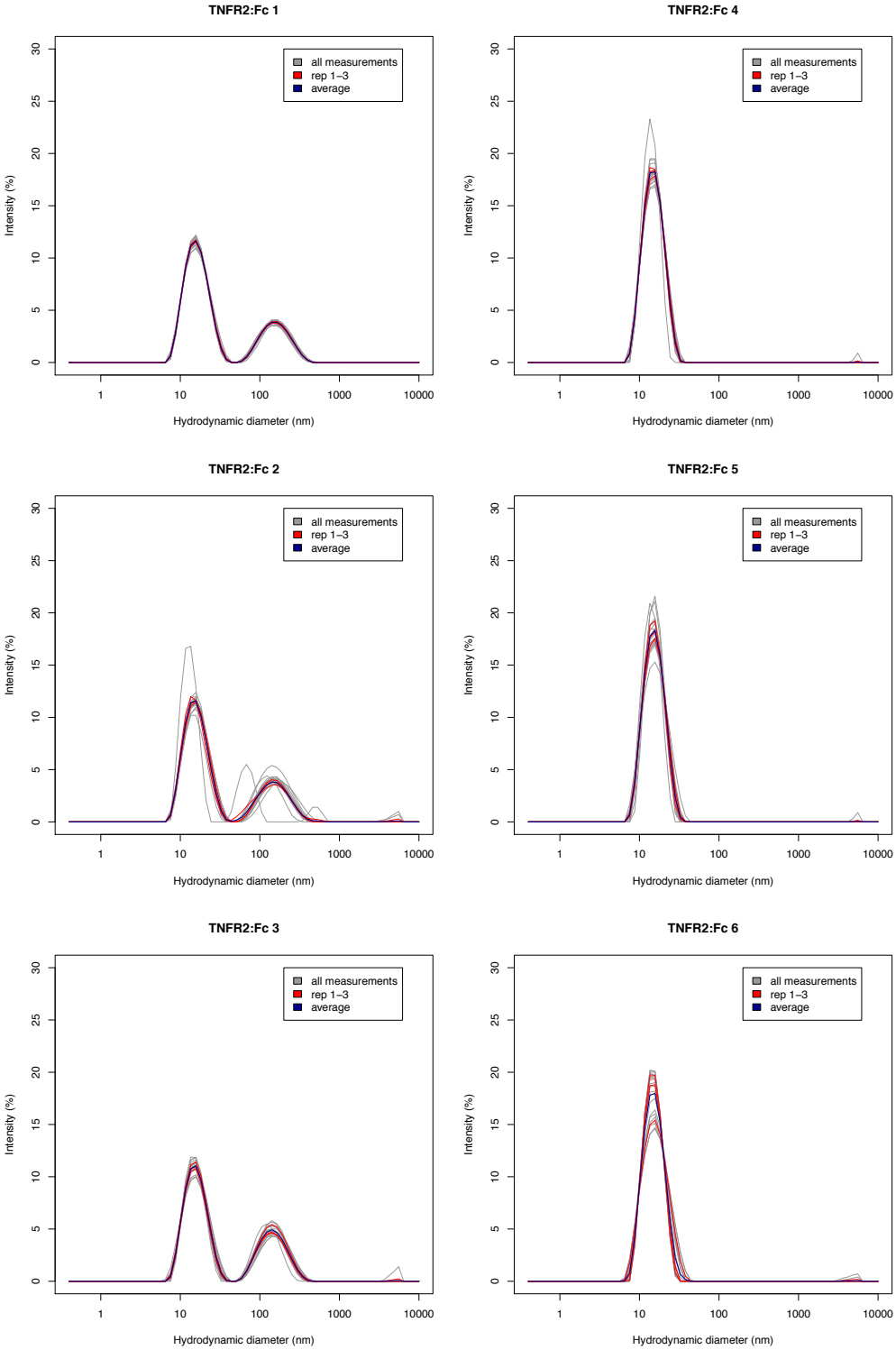


Figure B5: DLS measurements showing the particle size distributions (intensity values) of the six etanercept samples diluted to 1 mg/mL with PBS. The three biosimilars (samples 1–3) are presented on the left and the three innovator samples (4–6) are shown on the right. Samples were recorded in triplicate with six measurements for each replicate. All 18 measurements are shown in grey, averages of 6 measurements for each replicate are shown in red and total average in blue.

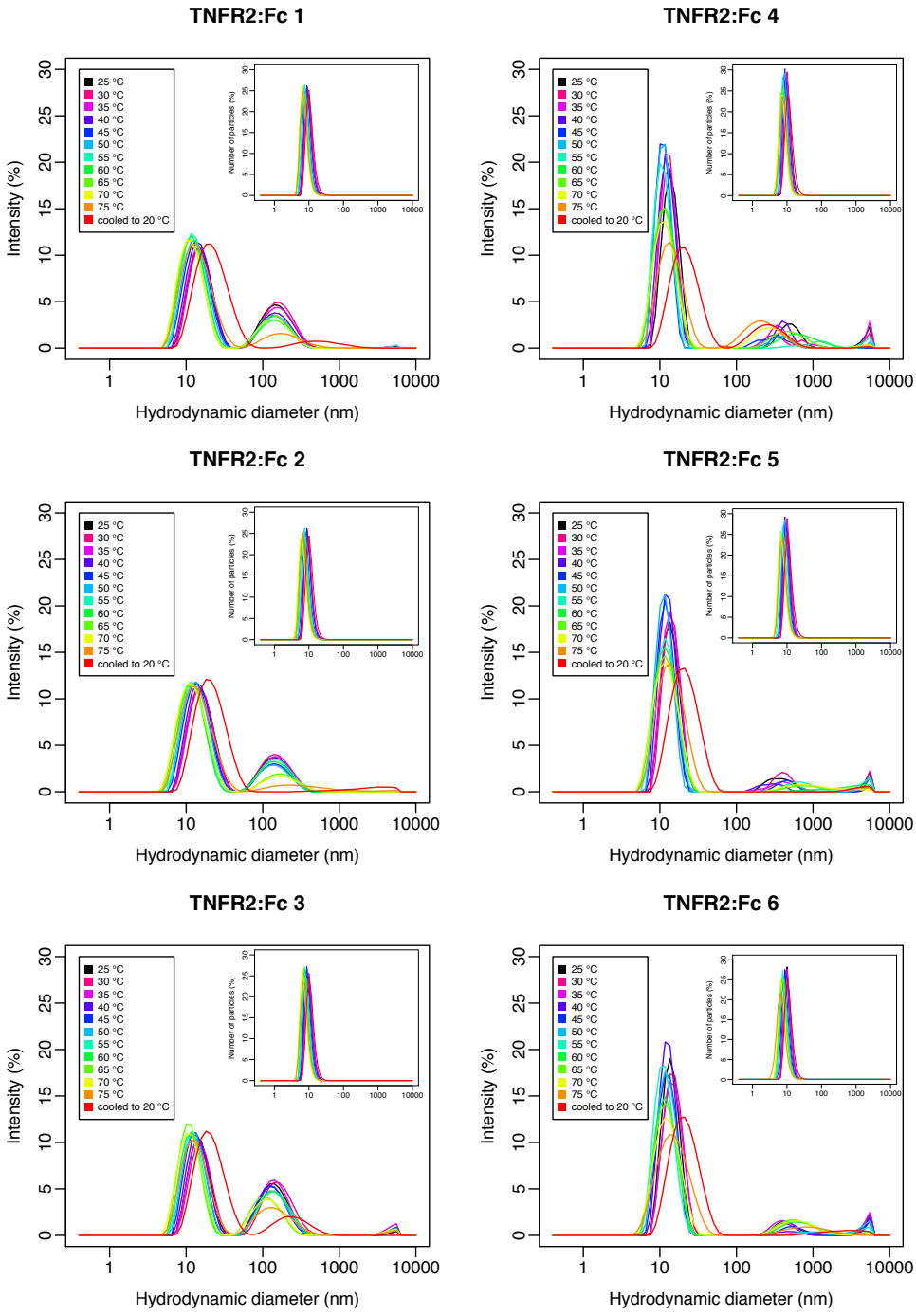


Figure B6: Size distributions (presented in intensity values, nm) for the six etanercept samples at a temperature range of 25–75°C and after the samples were cooled down to 20°C. The distribution becomes wider and shifts to larger values at the end of the experiment (in red the distribution after the samples were cooled down to 20 °C). The insets show the distributions in number values. The three biosimilars (samples 1–3) are presented on the left and the three innovator samples (4–6) are shown on the right. The presented data were averaged over 6 measurements.

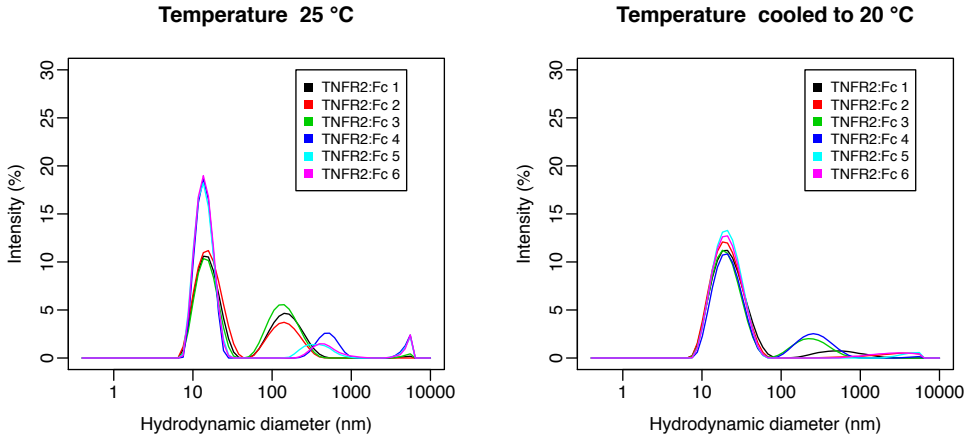


Figure B7: Size distributions presented in intensity values (nm) for the six samples at 25 °C (left), and after the samples were cooled down to 20 °C (right). The distributions become wider at the end of the experiment and the size of the protein shifts to larger values. Both observations indicate the appearance of larger particles. The presented data are averaged over 6 measurements.

Table B11: Overall mean particle sizes presented in terms of number values (nm) and the standard deviations (over six measurements) with a 95% confidence interval are shown for each sample at a temperature range of 25–75°C and after the samples were cooled down to 20°C.

Sample	25 °C	30 °C	35 °C	40 °C	45 °C	50 °C	55 °C	60 °C	65 °C	70 °C	75 °C	Cooled to 20 °C
TNFR2:Fc 1	10.0±0.2	10.4±0.2	9.9±0.4	9.5±0.2	8.7±0.2	8.5±0.3	8.2±0.2	7.7±0.1	7.5±0.1	7.2±0.2	7.5±0.2	11.0±0.5
TNFR2:Fc 2	9.8±0.2	10.0±0.3	9.7±0.4	9.5±0.3	8.7±0.3	8.4±0.2	8.1±0.3	7.6±0.2	7.0±0.2	6.7±0.2	7.3±0.3	11.0±0.5
TNFR2:Fc 3	10.3±0.3	10.5±0.8	10.4±0.6	9.8±0.2	9.3±0.3	8.7±0.2	8.4±0.2	8.1±0.2	7.6±0.4	7.2±0.2	7.6±0.5	11.3±0.6
TNFR2:Fc 4	10.8±0.2	10.4±0.4	10.3±0.4	9.8±0.4	9.0±0.2	9.0±0.2	8.3±0.2	8.0±0.3	7.8±0.1	7.7±0.5	8.4±0.8	11.9±0.6
TNFR2:Fc 5	10.4±0.3	10.9±0.1	10.4±0.3	9.8±0.0	9.3±0.1	9.0±0.1	8.4±0.1	7.7±0.2	7.5±0.1	7.4±0.3	7.9±0.2	11.6±0.3
TNFR2:Fc 6	10.6±0.3	10.8±0.2	10.6±0.2	9.8±0.1	9.2±0.3	8.7±0.1	8.2±0.3	7.9±0.2	8.0±0.3	7.4±0.2	7.0±0.8	11.6±0.2

CD Spectroscopy

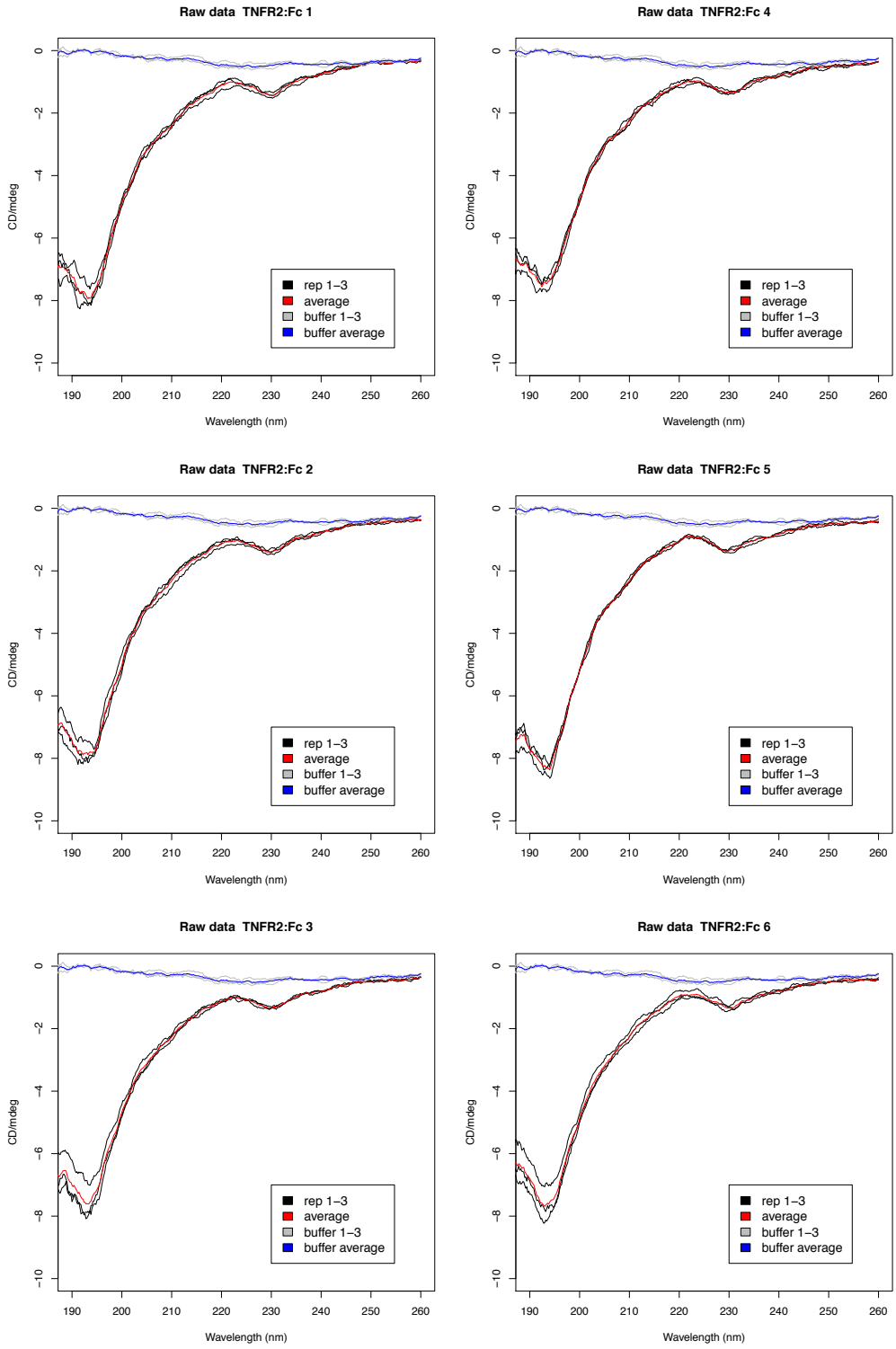


Figure B8: Raw far-UV CD data (mdeg) of the six etanercept samples and the reference buffer at the received concentrations (~0.9 mg/mL). The three biosimilars (samples 1–3) are presented on the left and the three innovator samples (4–6) are shown on the right. Samples and buffer were recorded in triplicate (black and grey respectively). The averaged data for the buffer and the samples are shown in blue and red respectively.

Table B12: Minima (wavelength and Δε values) of the CD spectra for the 6 etanercept samples in their original concentrations (~0.9 mg/mL) using a 0.01 cm demountable cuvette. The averaged values over three replicates are presented for each sample, as replicates appeared relatively noisy.

Sample	1st Min (nm)	Min value (Δε)	2nd Min (nm)	Min value (Δε)
TNFR2:Fc 1	193.2	-4.98	230.4	-0.58
TNFR2:Fc 2	192.2	-4.47	229.8	-0.48
TNFR2:Fc 3	192.8	-4.68	229.6	-0.48
TNFR2:Fc 4	192.4	-4.70	231.4	-0.49
TNFR2:Fc 5	194.0	-4.67	229.8	-0.44
TNFR2:Fc 6	193.0	-4.42	230.8	-0.44

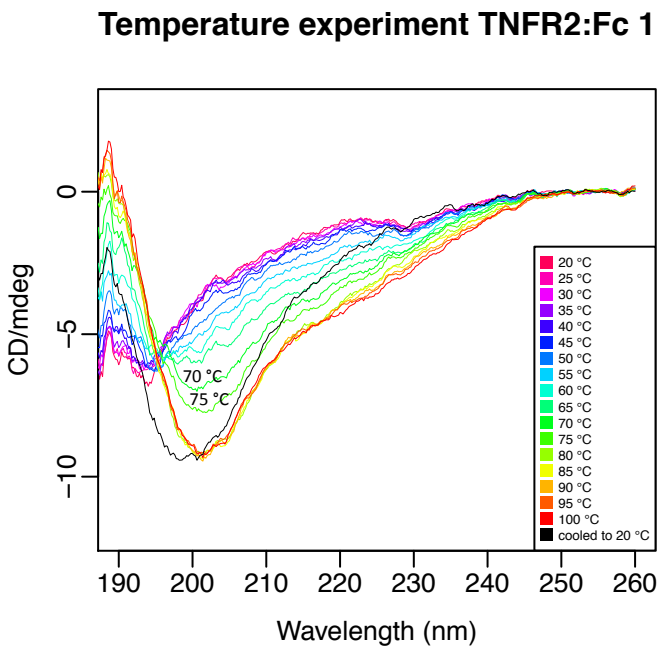


Figure B9: CD spectra of the diluted etanercept sample 1 at a temperature range of 20–100 °C, per 5 °C. The spectra were recorded at a concentration of 0.1 mg/mL, using a 0.1 cm fixed path length cuvette. The multiple crossing-points of the spectra at around 195 nm indicate that etanercept has more than one transition melting temperatures. The fully unfolded structure was observed at around 80 °C, while the spectrum after the sample was cooled down to 20 °C (black line) presented a somewhat different shape than the spectra at high temperatures (80–100 °C), slightly shifted towards a lower wavelength.

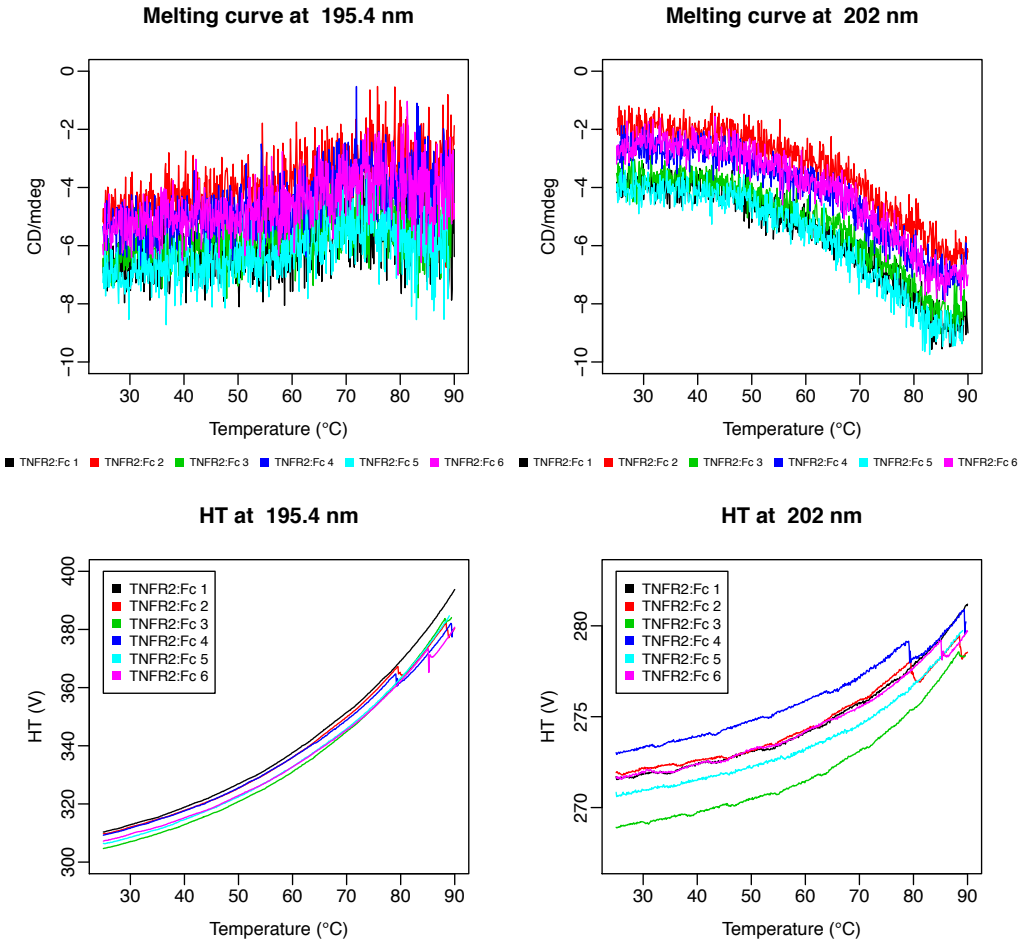


Figure B10: Melting curves recorded for the formulated etanercept samples that were diluted in two steps: to a concentration of 1 mg/mL with PBS, and a concentration of 0.1 mg/mL with water. The 0.1 cm fixed path length cuvette was used. Top: raw melting curves (mdeg) of the etanercept samples at 195.4 nm (left) and at 202.0 nm (right). Bottom: the respective HT (V) values, recorded at the same time as the melting curves, at 195.4 nm (left) and at 202.0 nm (right). The observed "dips" in the HT values for samples 2, 3, 4 and 6 can result from aggregated particles that settle at the bottom of the cuvette with time. However, the changes were very small, perhaps indicating that aggregation took place only to a limited extent.

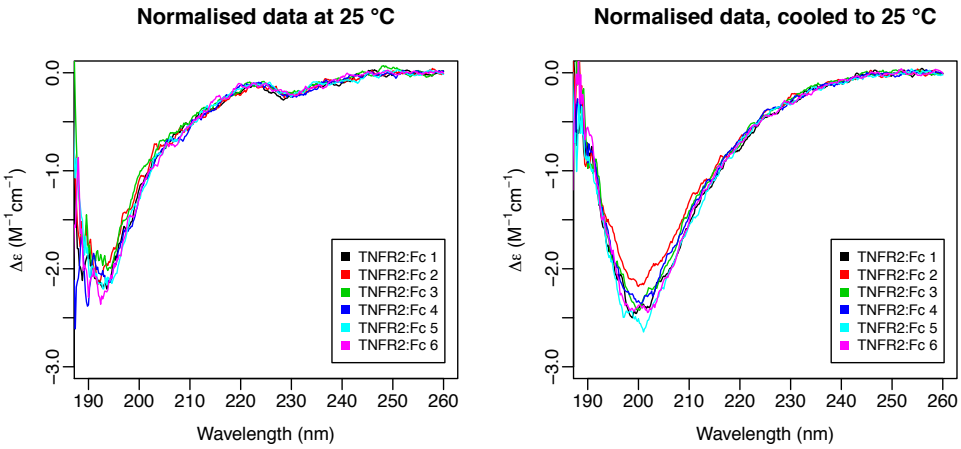


Figure B11: Comparison of the CD spectra of the diluted etanercept samples at a concentration of 0.1 mg/mL using a 0.1 cm fixed path length cuvette. Spectra were acquired before recording the melting curves (25 °C, left), and at the end of the temperature experiment after the samples were cooled down to 25 °C (right).

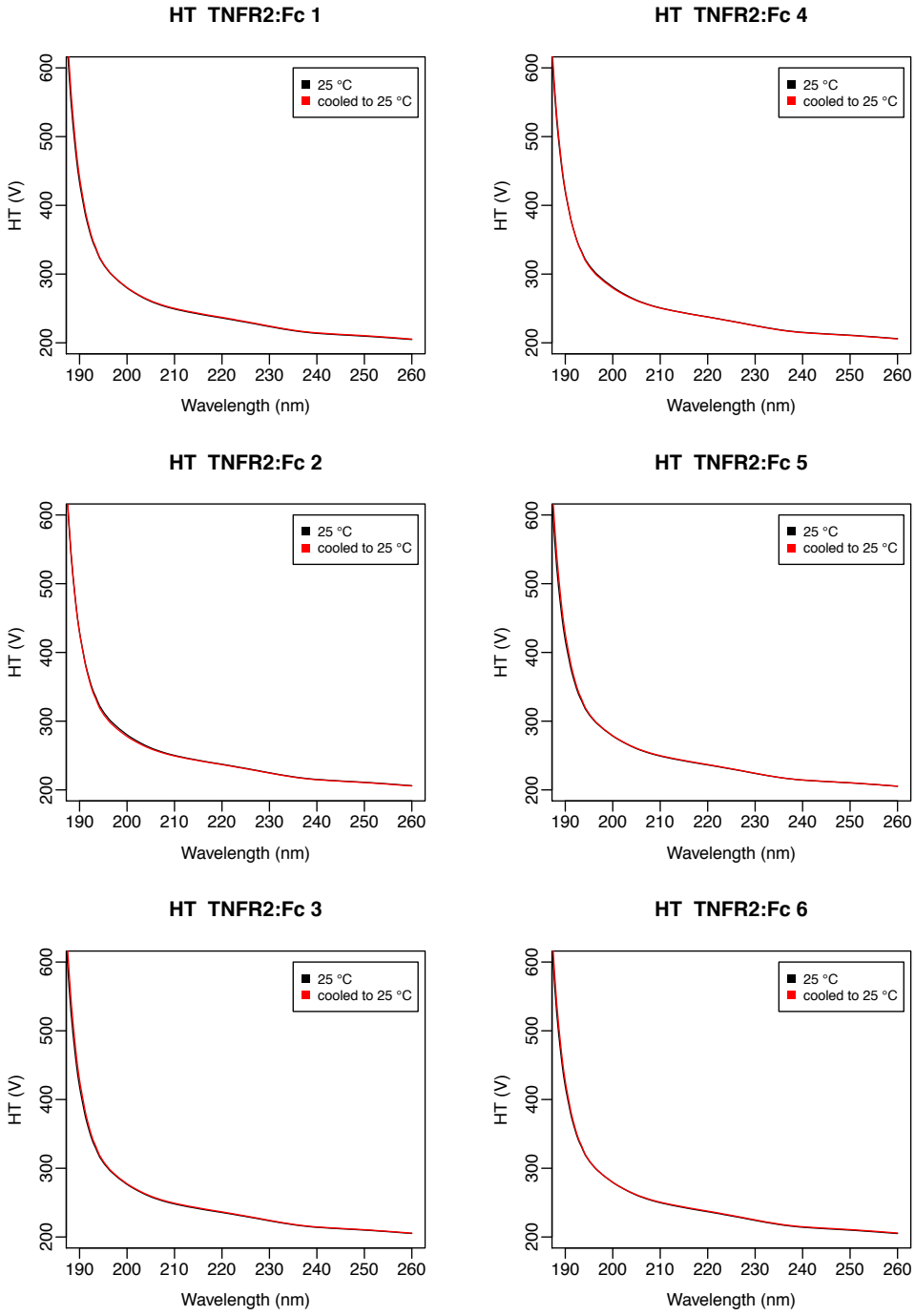


Figure B12: Comparison of the HT spectra of the diluted etanercept samples at a concentration of 0.1 mg/mL using a 0.1 cm fixed path length cuvette, recorded at the same time as the CD spectra from Figure B11. The HT spectra at 25 °C (before recording the melting curves, black) and at the end of the temperature experiment after the samples were cooled down to 25 °C (red) are identical, suggesting that no significant aggregation took place. The three biosimilar samples (1–3) are presented on the left and the three innovator samples (4–6) are shown on the right.

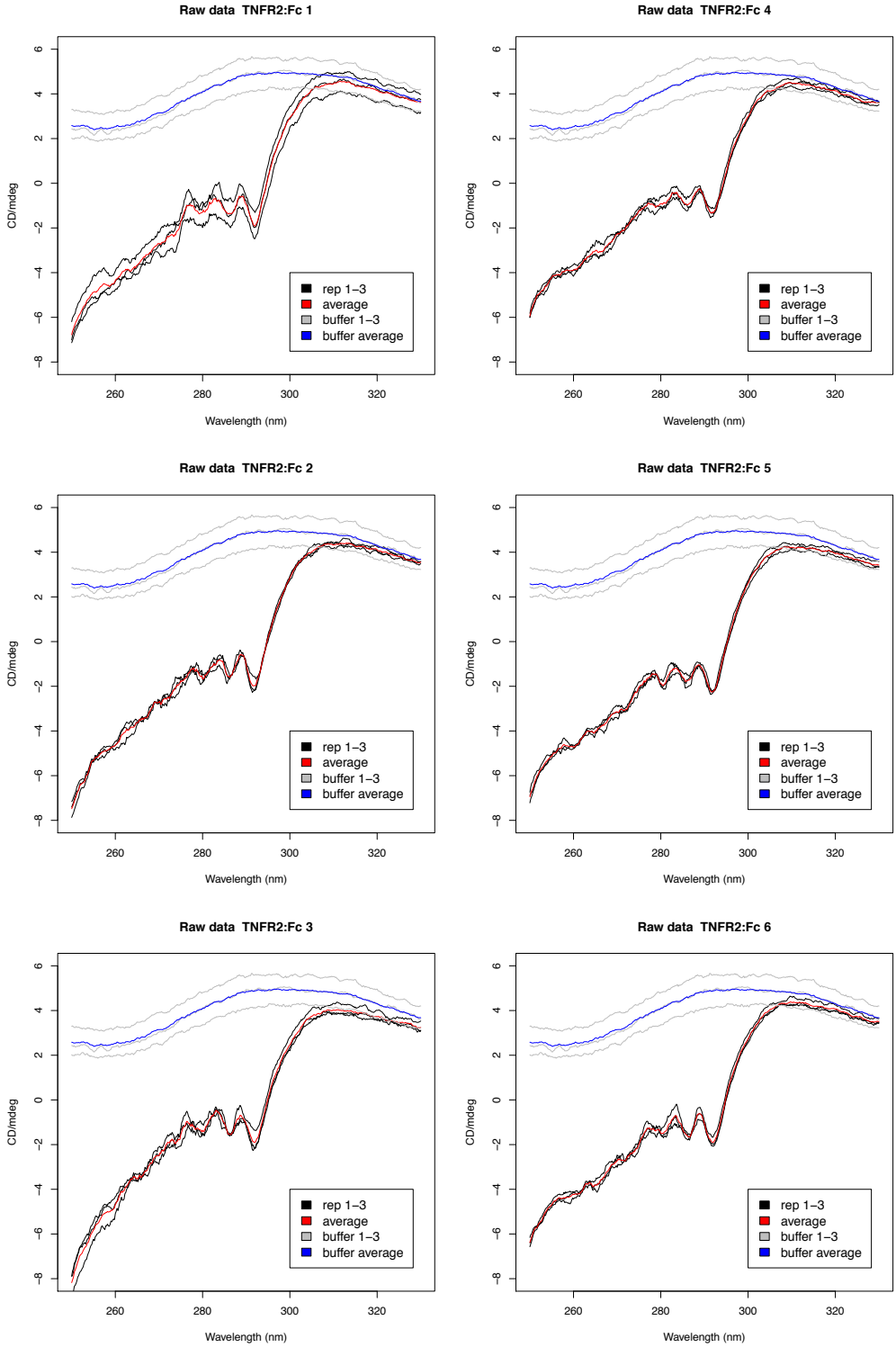


Figure B13: Raw near-UV CD data (mdeg) of the six etanercept samples and the reference buffer in the received concentrations (~0.9 mg/mL). The three biosimilars (samples 1–3) are presented on the left and the three innovator samples (4–6) are shown on the right. Samples and buffer were recorded in triplicate (black and grey respectively). The averaged data for the buffer and the samples are shown in blue and red respectively.

Table B13: Secondary structure estimates and NRMSD values from analysis of the CD spectra of etanercept. The results from CDSSTR, SELCON3 (using reference sets 1–7) and SSNN (reference set 7) showed a large variation due to the unique spectrum of the protein. Results that presented ≤10% deviation from the secondary structure that was calculated by combination of the PDB crystal structures are highlighted in pale green.

Sample	Program	Ref. set	Helix	Sheet	Turns	Other	NRMSD
TNFR2:Fc 1	CDSSTR	1	-0.02	0.26	0.31	0.46	0.098
	SELCON3		-0.03	0.14	0.30	0.59	0.214
	CDSSTR	2	0.05	0.25	0.24	0.46	0.102
	SELCON3		0.07	0.19	0.16	0.58	0.514
	CDSSTR	3	0.03	0.37	0.24	0.37	0.140
	SELCON3		0.01	0.26	0.26	0.47	0.453
	CDSSTR	4	0.03	0.35	0.27	0.35	0.222
	SELCON3		0.02	0.27	0.25	0.47	0.467
	CDSSTR	5	0.03	0.25	0.15	0.58	0.220
	SELCON3		0.02	0.17	0.10	0.71	0.608
	CDSSTR	6	0.00	0.23	0.14	0.63	0.165
	SELCON3		-0.02	0.14	0.26	0.61	0.347
	CDSSTR		0.02	0.18	0.14	0.66	0.075
	SELCON3	7	0.00	0.04	0.07	0.90	0.683
	SSNN		0.04	0.10	0.07	0.79	0.158
TNFR2:Fc 2	CDSSTR	1	-0.02	0.28	0.28	0.46	0.118
	SELCON3		-0.03	0.22	0.31	0.50	0.342
	CDSSTR	2	0.04	0.24	0.22	0.50	0.091
	SELCON3		0.06	0.30	0.17	0.47	0.450
	CDSSTR	3	0.02	0.35	0.24	0.38	0.171
	SELCON3		0.01	0.08	0.27	0.64	0.208
	CDSSTR	4	0.03	0.34	0.26	0.37	0.179
	SELCON3		0.11	0.33	0.21	0.35	0.204
	CDSSTR	5	0.02	0.26	0.15	0.57	0.235
	SELCON3		0.01	0.18	0.09	0.72	0.597
	CDSSTR	6	0.00	0.22	0.15	0.63	0.119
	SELCON3		-0.01	0.11	0.23	0.68	0.393
	CDSSTR		0.02	0.21	0.14	0.63	0.109
	SELCON3	7	0.02	0.21	0.18	0.59	0.700
	SSNN		0.05	0.11	0.09	0.76	0.160
TNFR2:Fc 3	CDSSTR	1	-0.02	0.24	0.30	0.48	0.123
	SELCON3		-0.02	0.16	0.30	0.56	0.274
	CDSSTR	2	0.05	0.20	0.26	0.48	0.069
	SELCON3		0.01	0.46	0.29	0.24	0.079
	CDSSTR	3	0.02	0.33	0.26	0.39	0.122
	SELCON3		0.01	0.05	0.27	0.67	0.159
	CDSSTR	4	0.03	0.33	0.26	0.39	0.168
	SELCON3		0.10	0.27	0.22	0.41	0.448
	CDSSTR	5	0.03	0.24	0.16	0.58	0.218
	SELCON3		0.01	0.18	0.10	0.72	0.591
	CDSSTR	6	0.00	0.27	0.15	0.58	0.138
	SELCON3		-0.01	0.11	0.24	0.67	0.353
	CDSSTR		0.02	0.21	0.15	0.62	0.104
	SELCON3	7	0.03	0.20	0.17	0.60	0.642
	SSNN		0.05	0.11	0.08	0.76	0.159

Sample	Program	Ref. set	Helix	Sheet	Turns	Other	NRMSD
TNFR2:Fc 4	CDSSTR	1	-0.03	0.30	0.33	0.40	0.131
	SELCON3		-0.02	0.14	0.30	0.57	0.250
	CDSSTR	2	0.05	0.22	0.22	0.51	0.064
	SELCON3		0.01	0.45	0.30	0.25	0.074
	CDSSTR	3	0.02	0.37	0.24	0.37	0.164
	SELCON3		0.00	0.06	0.28	0.67	0.168
	CDSSTR	4	0.03	0.35	0.27	0.36	0.158
	SELCON3		0.10	0.24	0.20	0.46	0.361
	CDSSTR	5	0.02	0.23	0.15	0.60	0.222
	SELCON3		0.01	0.18	0.10	0.72	0.598
	CDSSTR	6	0.01	0.25	0.15	0.59	0.123
	SELCON3		-0.01	0.11	0.23	0.67	0.360
	CDSSTR		0.02	0.19	0.13	0.66	0.097
	SELCON3	7	0.03	0.20	0.17	0.60	0.664
	SSNN		0.05	0.11	0.08	0.76	0.158
TNFR2:Fc 5	CDSSTR	1	-0.03	0.25	0.29	0.49	0.142
	SELCON3		-0.02	0.13	0.30	0.59	0.216
	CDSSTR	2	0.04	0.21	0.23	0.52	0.071
	SELCON3		0.02	0.46	0.29	0.24	0.087
	CDSSTR	3	0.02	0.36	0.24	0.38	0.147
	SELCON3		0.00	0.10	0.29	0.61	0.211
	CDSSTR	4	0.02	0.35	0.26	0.37	0.143
	SELCON3		0.10	0.27	0.22	0.41	0.488
	CDSSTR	5	0.02	0.26	0.15	0.57	0.251
	SELCON3		0.00	0.19	0.10	0.72	0.593
	CDSSTR	6	0.00	0.26	0.18	0.56	0.131
	SELCON3		-0.01	0.10	0.23	0.68	0.344
	CDSSTR		0.01	0.20	0.14	0.64	0.109
	SELCON3	7	0.02	0.19	0.16	0.63	0.670
	SSNN		0.05	0.11	0.08	0.76	0.161
TNFR2:Fc 6	CDSSTR	1	-0.02	0.23	0.29	0.51	0.109
	SELCON3		-0.03	0.11	0.29	0.63	0.178
	CDSSTR	2	0.04	0.22	0.22	0.53	0.072
	SELCON3		0.10	0.22	0.15	0.53	0.451
	CDSSTR	3	0.02	0.34	0.25	0.39	0.185
	SELCON3		0.04	0.11	0.27	0.58	0.170
	CDSSTR	4	0.03	0.33	0.26	0.38	0.165
	SELCON3		0.12	0.28	0.21	0.39	0.292
	CDSSTR	5	0.02	0.25	0.15	0.58	0.232
	SELCON3		0.01	0.19	0.10	0.70	0.622
	CDSSTR	6	0.00	0.25	0.16	0.59	0.127
	SELCON3		-0.01	0.08	0.22	0.71	0.314
	CDSSTR		0.01	0.23	0.15	0.62	0.118
	SELCON3	7	0.03	0.15	0.13	0.69	0.410
	SSNN		0.05	0.11	0.09	0.75	0.155

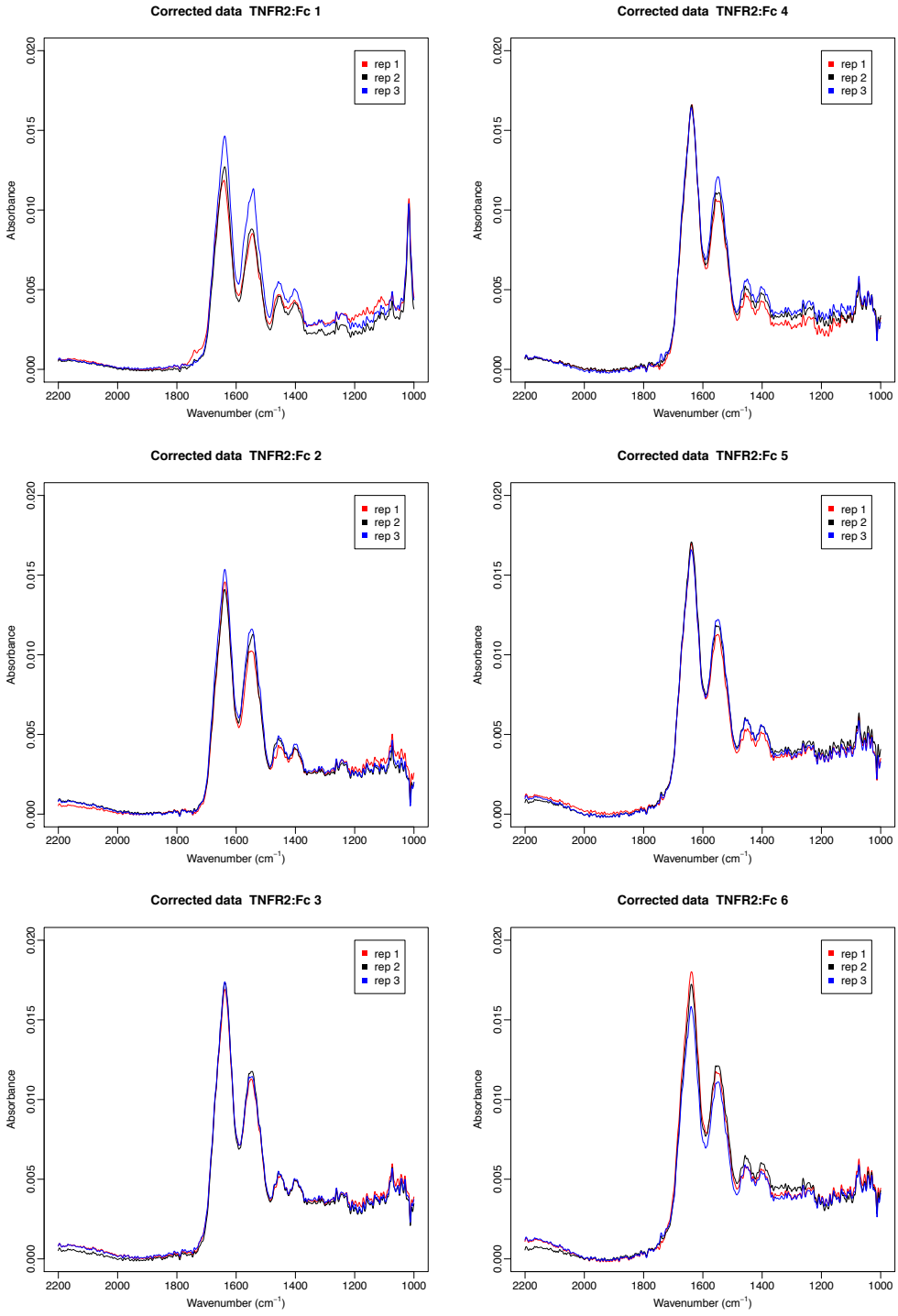


Figure B14: FT-IR data (absorbance units) of triplicate measurement of the etanercept samples (~0.9 mg/mL) following baseline subtraction and zeroing. The three biosimilars (samples 1–3) are presented on the left and the three innovator samples (4–6) are shown on the right.

Fluorescence

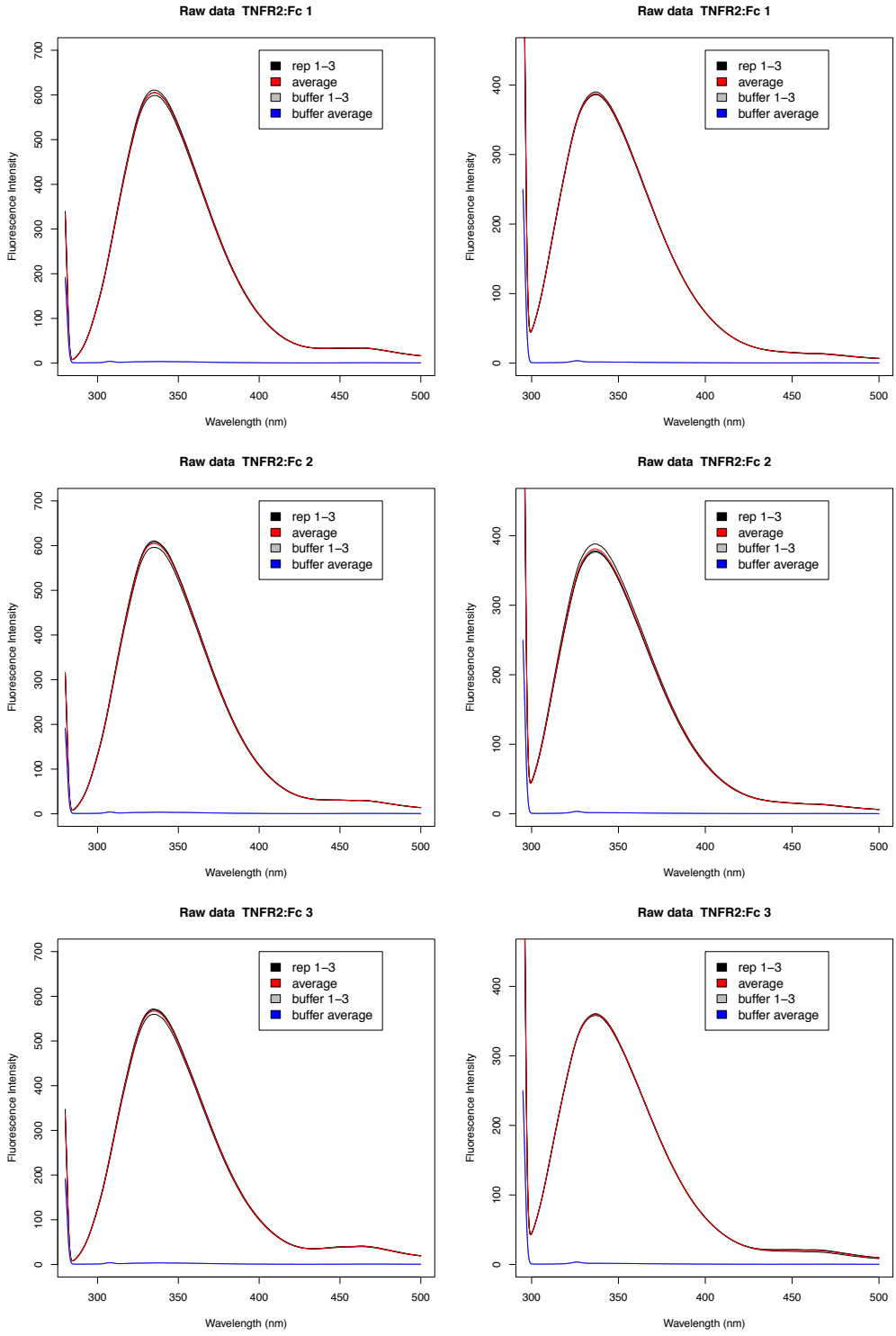


Figure B15: Raw fluorescence data (1st analysis) of the *biosimilar* etanercept samples (samples 1–3) and the reference buffer (ammonium bicarbonate) at the received concentrations (~0.9 mg/mL). Data for excitation wavelengths 280 nm (left) and 295 nm (right) are shown. Samples and buffer were recorded in triplicate (black and grey respectively). The averaged data for the buffer and the samples are shown in blue and red respectively.

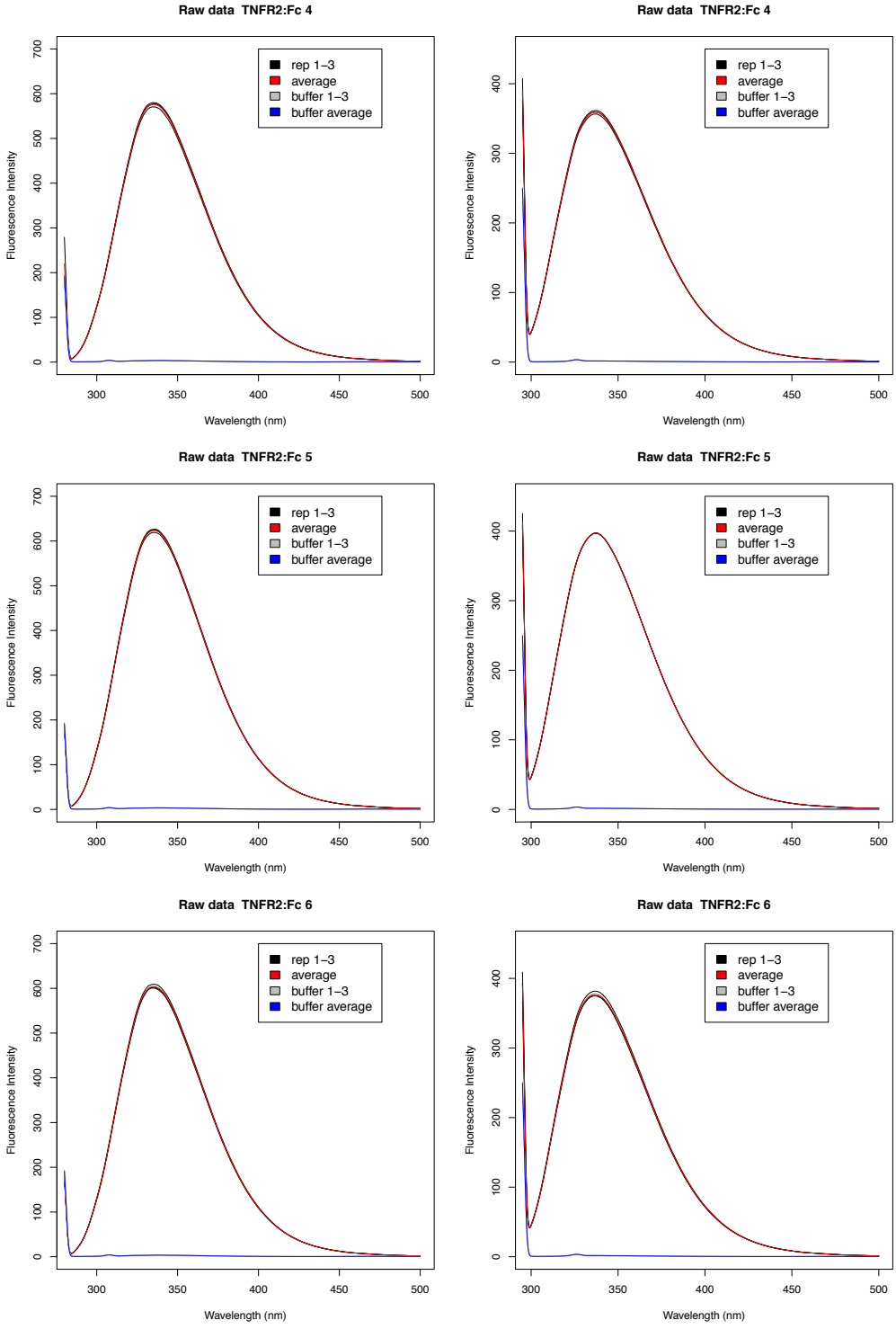


Figure B16: Raw fluorescence data (1st analysis) of the *innovator* etanercept samples (samples 4–6) and the reference buffer (ammonium bicarbonate) at the received concentrations (~0.9 mg/mL). Data for excitation wavelengths 280 nm (left) and 295 nm (right) are shown. Samples and buffer were recorded in triplicate (black and grey respectively). The averaged data for the buffer and the samples are shown in blue and red respectively.

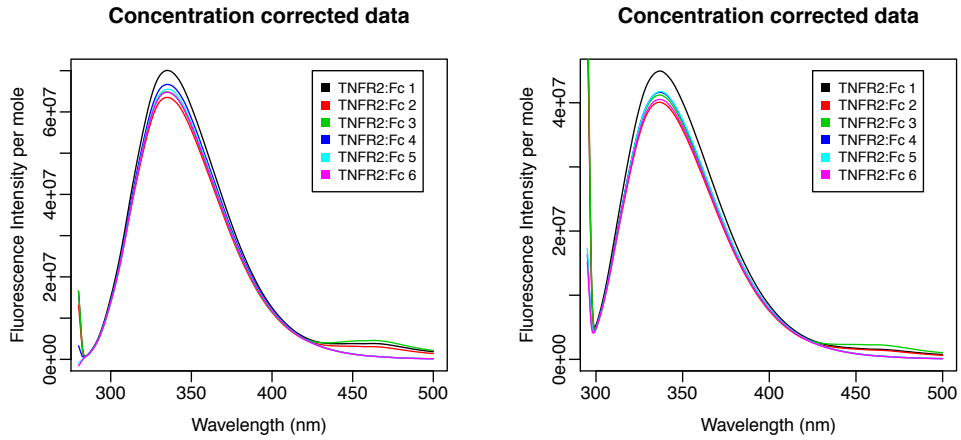


Figure B17: Averaged fluorescence emission spectra of the etanercept samples for excitation wavelengths 280 nm (left) and 295 nm (right), after correcting for the concentration variation among samples. The data are from the 1st analysis of the samples in ammonium bicarbonate buffer, at the received concentrations (~0.9 mg/mL).

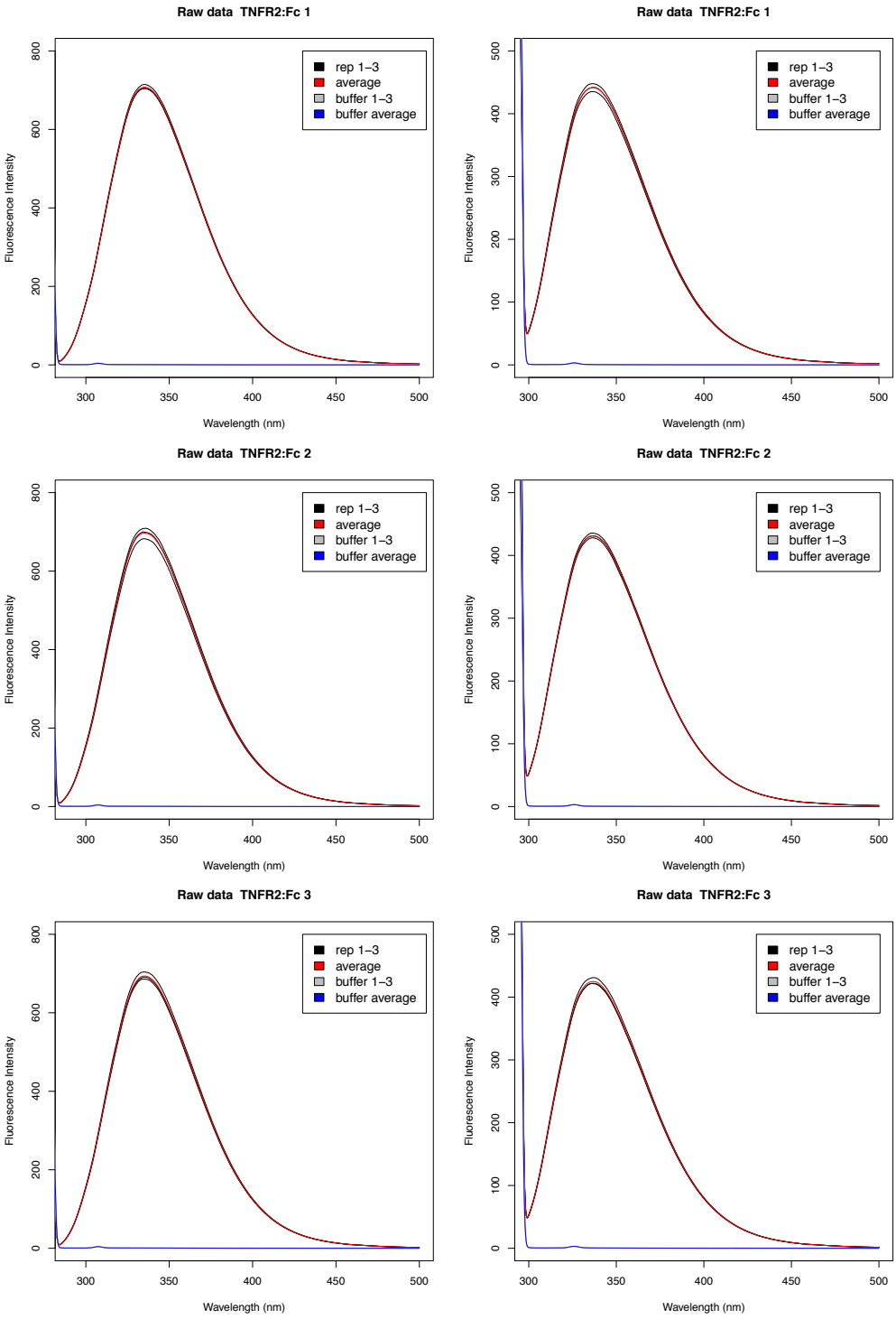


Figure B18: Raw fluorescence data (2nd analysis) of the *biosimilar* etanercept samples (samples 1–3) diluted to a concentration of 1 mg/mL and the reference buffer (PBS). Data for excitation wavelengths 280 nm (left) and 295 nm (right) are shown. Samples and buffer were recorded in triplicate (black and grey respectively). The averaged data for the buffer and the samples are shown in blue and red respectively.

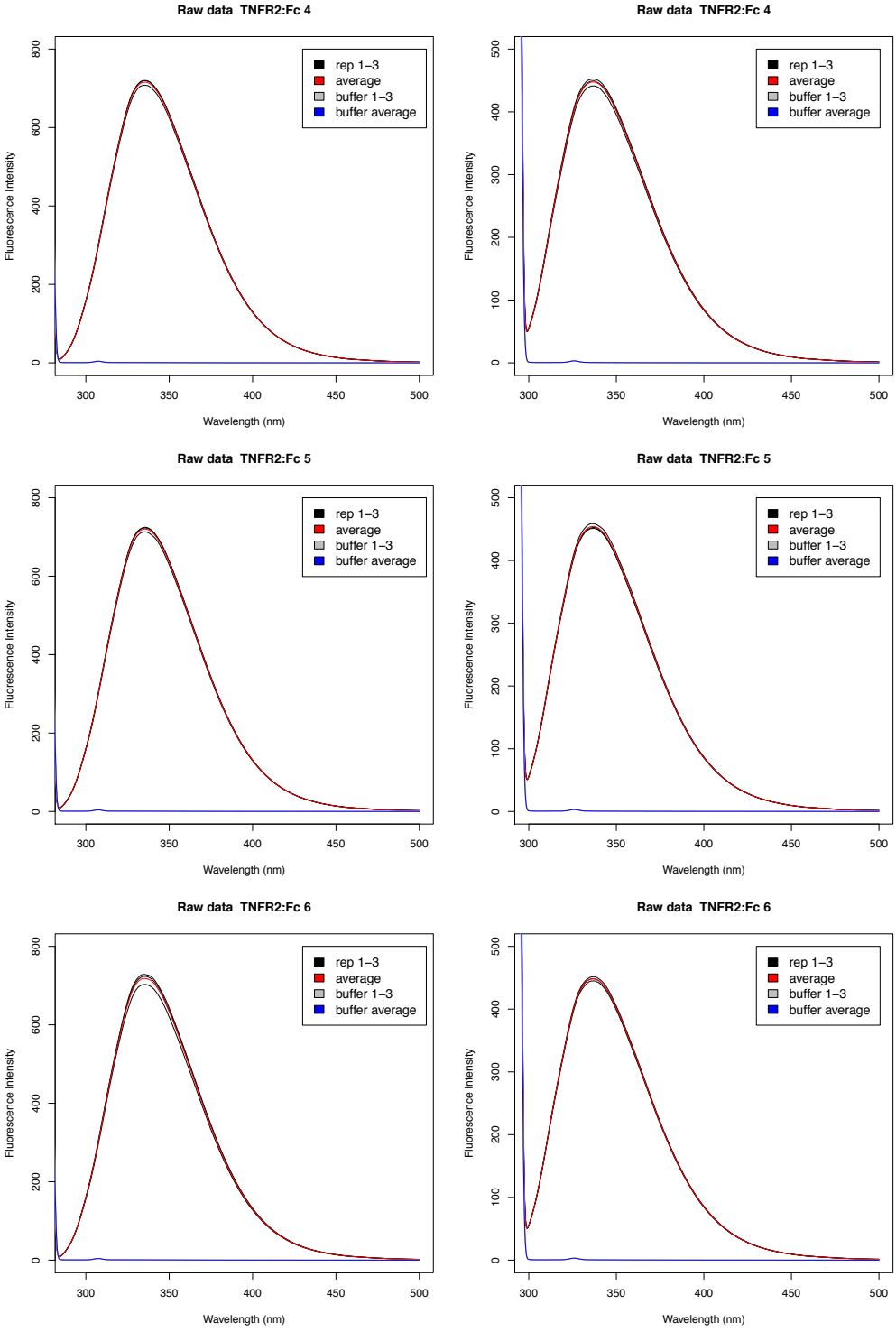


Figure B19: Raw fluorescence data (2nd analysis) of the *innovator* etanercept samples (samples 4–6) diluted to a concentration of 1 mg/mL and the reference buffer (PBS). Data for excitation wavelengths 280 nm (left) and 295 nm (right) are shown. Samples and buffer were recorded in triplicate (black and grey respectively). The averaged data for the buffer and the samples are shown in blue and red respectively.

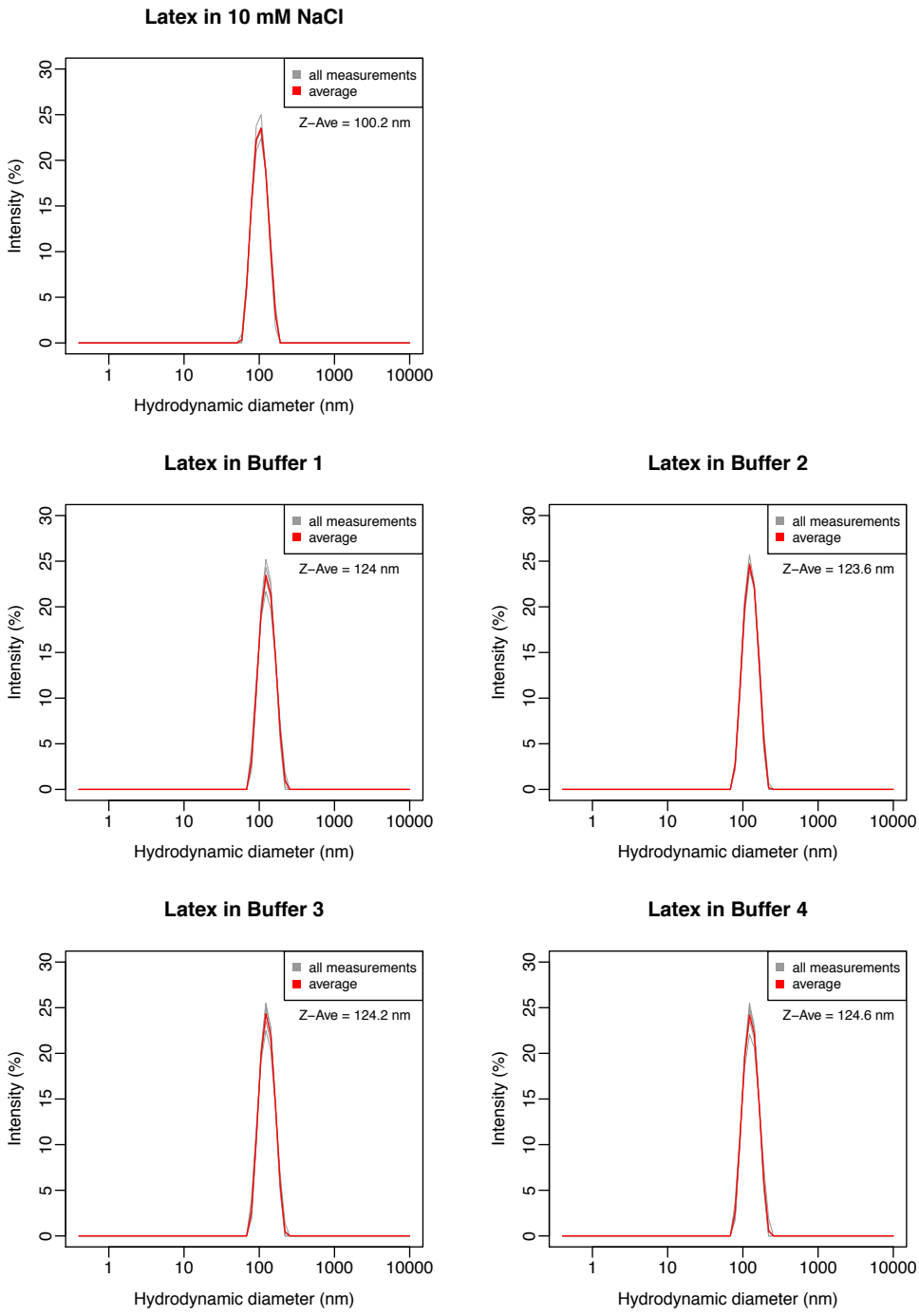


Figure B20: Top: Particle distribution by intensity of 100 nm latex spheres measured in 10 mM NaCl solution. Middle/Bottom: Particle distribution by intensity of 100 nm latex spheres measured in each one of the four buffers, prepared according to dilution of the respective samples. Individual measurements are shown in grey and average over 5 measurements is shown in red. The Z-Average values as calculated by the software are given, as an average over 5 measurements.

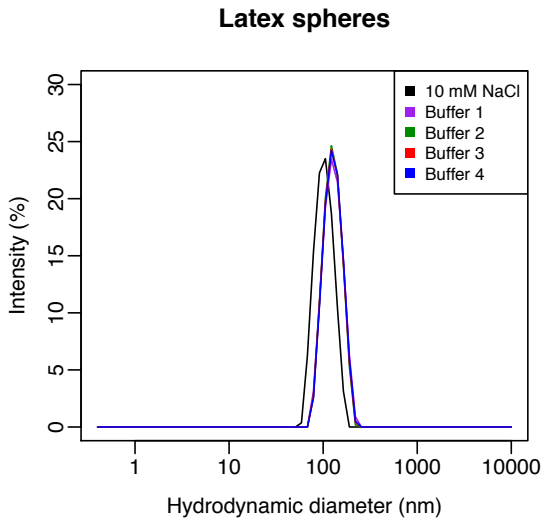


Figure B21: Average particle distributions by intensity of 100 nm latex spheres in NaCl (black) and in each one of the buffer 1–4 (purple, green, red and blue) solutions, prepared according to the dilution of the respective samples. All data are averaged over 5 measurements.

DLS Spectroscopy

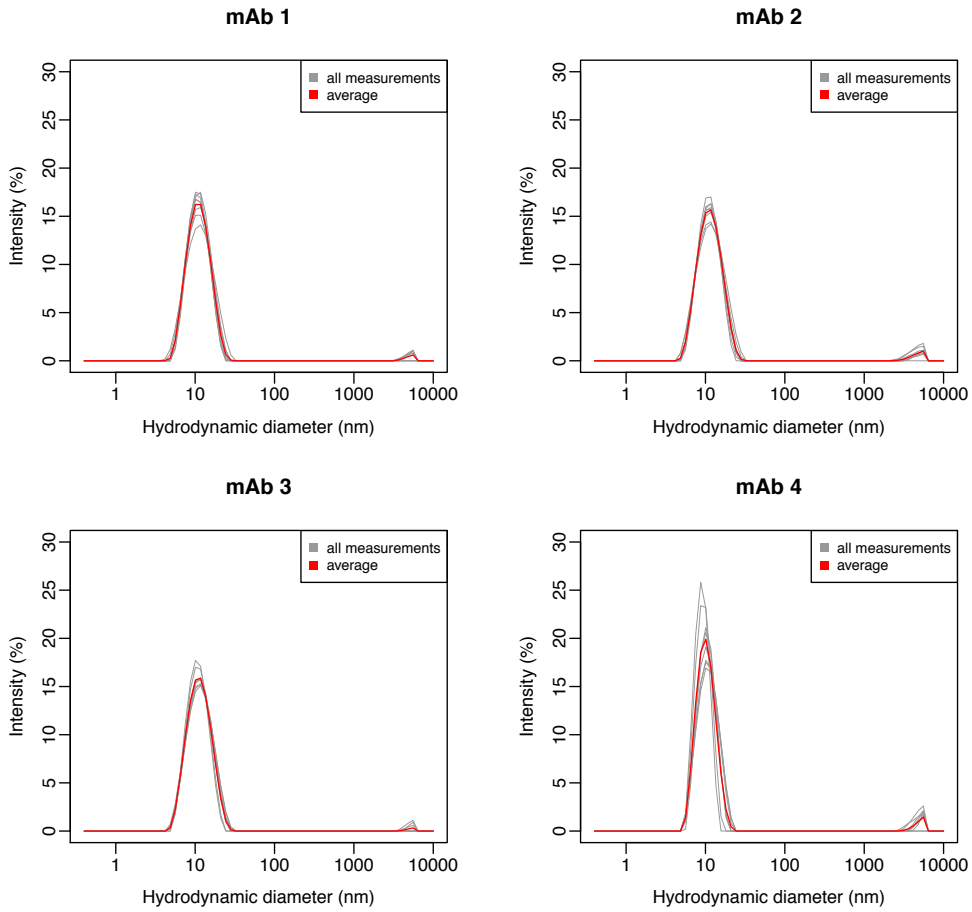


Figure B22: DLS measurements showing the particle size distributions (intensity values) of the four mAb samples diluted to 5 mg/mL with the provided formulation buffer (omitting Tween 80). All 8 measurements are shown in grey, average over 8 measurements is shown in red.

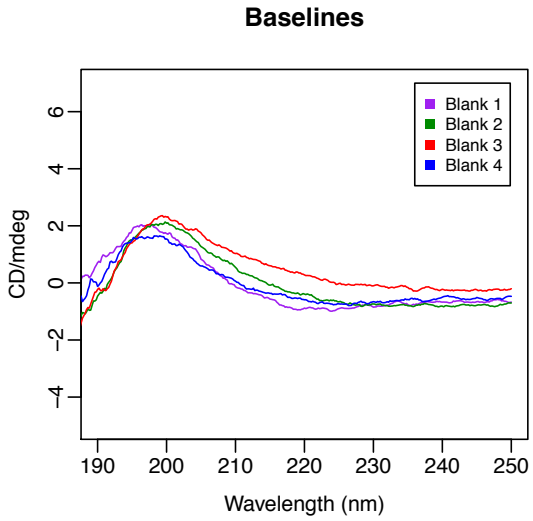


Figure B23: Recorded baselines before each one of the monoclonal antibody samples, after 1000-fold dilution of the original formulation with water and 100 mM sodium phosphate buffer, resulting in a final solution of 10 mM sodium phosphate buffer, 50 μM citric acid, 110 μM L-arginine, 240 μM L-arginine HCl and 0.00005% w/v Tween 80. A 0.1 cm fixed path length cuvette was used.

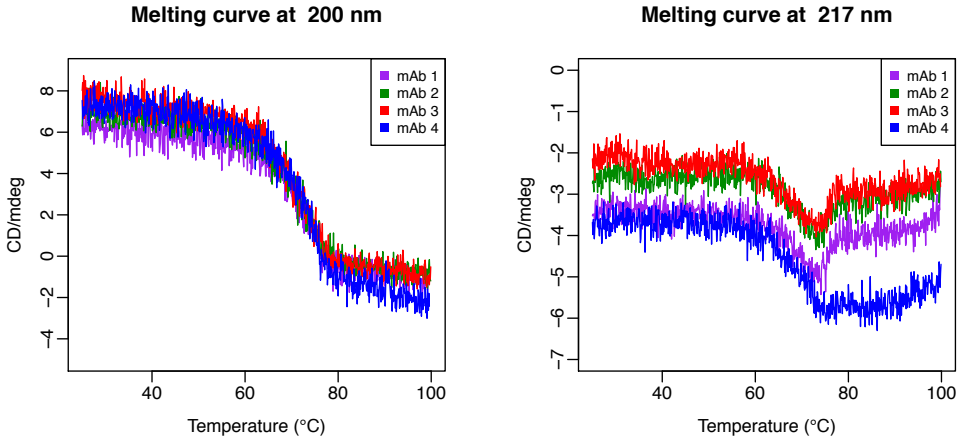


Figure B24: Raw melting curves (mdeg) of the monoclonal antibody samples at 200.0 nm (left) and 217.0 nm (right). The curves were recorded for the diluted with sodium phosphate buffer samples, resulting to a final solution of ~0.1 mg/mL protein in 10 mM sodium phosphate buffer. The 0.1 cm fixed path length cuvette was used.

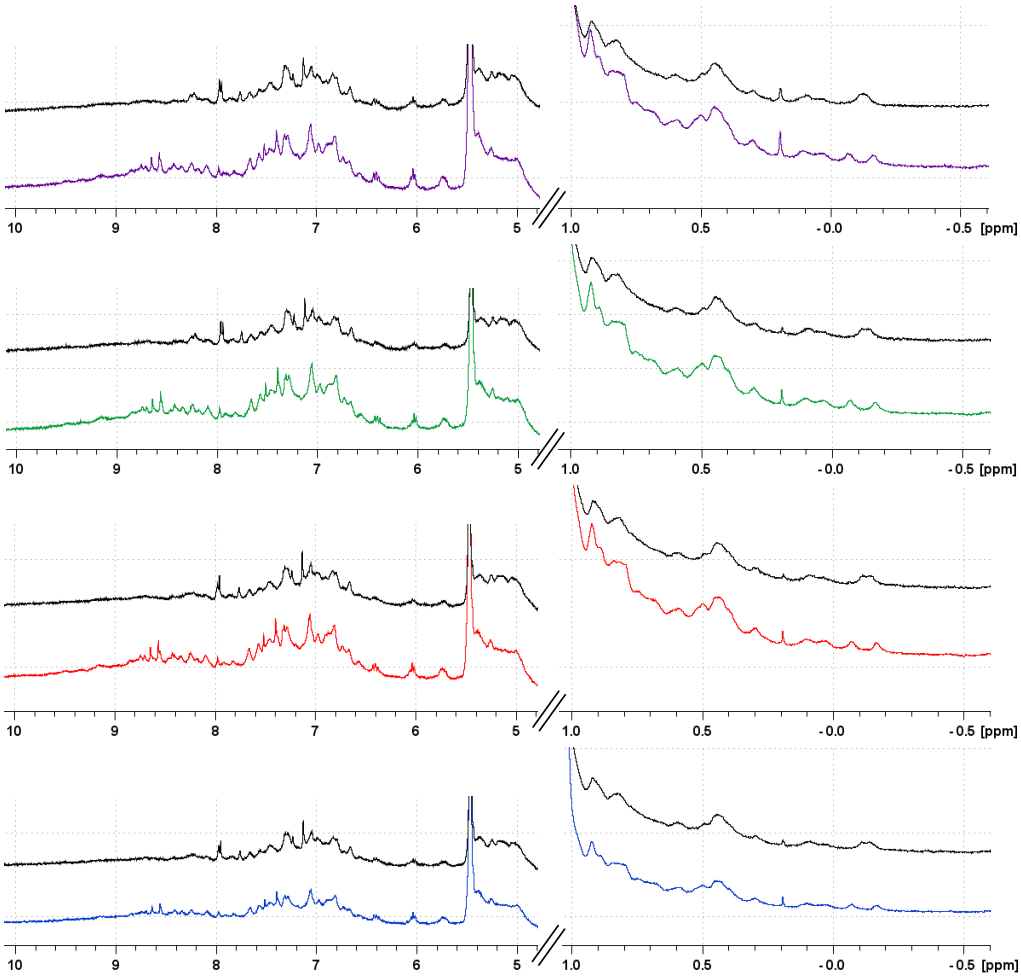


Figure B25: Comparison of the 1D ^{1}H-NMR spectra of the four monoclonal antibodies in sodium phosphate buffer at pH 5.4 (in colour) and pH 7.0 (in black). From top to bottom: mAb 1 is shown in purple (pH 5.4) and black (pH 7.0), mAb 2 is shown in green (pH 5.4) and black (pH 7.0), mAb 3 is shown in red (pH 5.4) and black (pH 7.0) and mAb 4 is shown in blue (pH 5.4) and black (pH 7.0).

Appendix C

THE UNIVERSITY OF WARWICK

Standard Operating Procedure

Department of Chemistry
University of Warwick

Near-UV and far-UV circular dichroism measurement of protein samples with instrument calibration and secondary structure determination

Document Reference:	SOP48.3
Prepared by:	Claire Broughton
Approved by:	Alison Rodger
Reason for issue:	
Date of first issue:	30th October 2002
Effective from:	22nd March 2016
Review date:	22nd March 2017

1. Document history

Document	Date of Issue	Reason for Issue
SOP48.1	30th October 2002	First Issue
SOP48.2	7th January 2013	Inclusion of Na[Co(EDDS)] calibration
SOP48.3	22nd March 2016	Inclusion of Benzene and Ammonia solution calibration

2. Objective

The objective of this Standard Operating Procedure (SOP) is to describe the calibration of the Jasco instrument and the measurement and analysis of near-UV and far-UV circular dichroism (CD) spectra of proteins.

3. Introduction

This SOP details the procedure, with reference to existing SOP's, for analysing the secondary structure of the protein samples using circular dichroism spectroscopy and Curtis Johnson's fitting program CDsstr as downloaded, saved to disk and modified in our laboratory for data collected from 260 nm to 190 nm or 195 nm. The program (CDsstr) for the SOP has been saved and archived for this purpose and will not be updated with any new versions that are released. It also contains instructions for running near UV CD spectra if required.

4. Safety

A full risk assessment must be completed prior to the delivery of the samples.

5. Responsibility

Any group member trained in the implementation of this SOP may carry out the required work.

6. Associated Documents

None

7. Equipment

Jasco V-660 UV/Vis spectrometer
Jasco J-720 Spectropolarimeter
Jasco J-815 Spectropolarimeter
Jasco J-1500 Spectropolarimeter
Mettler Toledo XP2U balance
Oxford A1204 Analytical balance
Sartorius S4 balance
CDsstr Original Version
0.01 mm rectangular demountable cuvette (Starna Ltd)
0.1 mm rectangular demountable cuvette (Starna Ltd)
1 mm rectangular cuvette (Starna Ltd)
1 cm rectangular cuvette (Starna Ltd)

8. Materials

Potassium Hydroxide AnalaR BDH 102104V
Potassium Chromate Aldrich 21,661-5, 99% ACS reagent
Ammonium d-camphor-10-sulphonate Katayama Chemical Co 05-1251
Na[Co(S,S-EDDS)] and Na[Co(R,R-EDDS)] University of Warwick
Ammonia solution 35% FSA, Analytical Reagent A/3280/PB17 Batch 22044111
Benzene Aldrich 40,176-5, 99.8%
Ethanol 'absolute' AnalaR BDH 10107EP
De-ionised water 18.2 MΩ Elga MaximaMillipore MilliQ
Buffers as required
Protein samples as supplied

9. Procedure

1. On receipt store the samples at 5 ± 3°C or as specified by the supplier.

2. *To assign the path length of the demountable 0. 01 mm cuvette proceed as follows:*
   2.1 Prepare a 0.2 M potassium chromate solution by accurately weighing 0.971g potassium chromate and transferring it to a 25 mL volumetric flask. Add approximately 20 mL water and 1 pellet of potassium hydroxide. Make up to volume. Mix well.
   2.2 Fill the demountable cell with the potassium chromate solution and measure the UV/vis absorption spectrum following the SOP for the spectrometer used. Note: it is important to always assemble the cell the same way. Calculate the pathlength of the cell using the Beer Lambert law $A=\varepsilon c l$ where A is the absorbance, c is the exact concentration of potassium chromate solution in moles dm^{-3}, ε is the extinction coefficient (4830 mol^{-1} dm^3 cm^{-1} at 372 nm), and *l* is the pathlength in cm.
   2.3 Repeat step 2.2 twice. Calculate the average value and record this as the nominal path length of the cuvette.

3. *To assign the pathlength of the demountable 0.1 mm cuvette proceed as follows:*
   3.1 Prepare a 0.02 M potassium chromate solution by accurately weighing 0.0971g potassium chromate and transferring it to a 25 mL volumetric flask. Add approximately 20 mL water and 1 pellet of potassium hydroxide. Make up to volume. Mix well. Alternatively accurately dilute the solution made in step 2 by a factor of 10.

3.2 Fill the demountable cell with the potassium chromate solution and measure the UV/vis absorption spectrum following the SOP for the spectrometer used. Note: it is important to always assemble the cell the same way. Calculate the pathlength of the cell using the Beer Lambert law $A=\varepsilon c l$ where A is the absorbance, c is the exact concentration of potassium chromate solution in moles dm^{-3}, ε is the extinction coefficient (4830 mol^{-1} dm^3 cm^{-1} at 372 nm), and *l* is the pathlength in cm.

3.3 Repeat step 3.2 twice. Calculate the average value and record this as the nominal path length of the cuvette.

4. *To calibrate the wavelength accuracy of the spectropolarimeter proceed as follows:*

4.1 Create a new folder for the current date within the instrument folder 01 Calibration on the network drive.

4.2 After the instrument has been on for more than 30 minutes measure the spectrum of the neodymium filter with the following parameters:

Channels:	CD/HT/Abs
DIT:	0.25 s
Band width:	1.0 nm
Wavelength range:	610 nm to 560 nm
Data pitch:	0.025 nm
Scan speed:	20 nm/min
Accumulations:	1

4.3 Save the file as “01 Neodymium Filter.jws”. Note the maximum HT value for the peak and the wavelength at which the HT is maximal, which should be 586 ± 0.8 nm. Select Processing > Common Options > Channel Conversion. Select Channel 2, click Add > OK. Select Processing > Peak Processing > Peak Find. Save the file as “01 Neodymium Filter Peaks.jws”. If the wavelength accuracy is not within specification, ensure the shift is constant across the wavelength range by checking with ACS as below. If the shift is constant recalibrate the spectrum accordingly. If it is not consult Professor Rodger who will call in an engineer.

4.4 Measure the spectrum of the holmium filter with the following parameters:

Channels:	CD/HT/Abs
DIT:	0.125 s
Band width:	1.0 nm
Wavelength range:	700 nm to 230 nm
Data pitch:	0.1 nm
Scan speed:	200 nm/min
Accumulations:	1

4.5 Save the file as “02 Holmium Filter.jws”. Note the maximum HT value for the peaks and compare to the manufacturer notes. The wavelength at which the HT is maximal should be within ± 0.8 nm of the reported value. Select Processing > Common Options > Channel Conversion. Select Channel 2, click Add > OK. Select Processing > Peak Processing > Peak Find. Save the file as “01 Holmium Filter Peaks.jws”. If the wavelength accuracy is not within specification, ensure the shift is constant across the wavelength range by checking with ACS as below. If the shift is constant recalibrate the spectrum accordingly. If it is not consult Professor Rodger who will call in an engineer.

4.6 Place 100 uL benzene in a 1 cm cuvette. Leave open in the fume hood for 10 mins, seal with a suitable lid and measure the benzene vapour spectrum with the following parameters:

Channels:	CD/HT/Abs
DIT:	0.125 s
Band width:	0.2 nm
Wavelength range:	265 nm to 240 nm
Data pitch:	0.05 nm
Scan speed:	50 nm/min
Accumulations:	1

4.7 Save the file as “03 Benzene Vapour.jws”. Note the maximum HT value for the peaks. The HT maxima should be at 259 nm, 253 nm, 247.2 nm and 241.7 nm within 0.8 nm. Select

Processing > Common Options > Channel Conversion. Select Channel 2, click Add > OK. Select Processing > Peak Processing > Peak Find. Save the file as “03 Benzene Vapour Peaks.jws”. If the wavelength accuracy is not within specification, ensure the shift is constant across the wavelength range by checking with ACS as below. If the shift is constant recalibrate the spectrum accordingly. If it is not consult Professor Rodger who will call in an engineer.

4.8 Place 10 uL ammonia solution in a 1 mm cuvette. Leave open in the fume hood for 10 mins. Seal the cuvetter with a suitable lid and measure the ammonia vapour spectrum with the following parameters:

Channels:	CD/HT/Abs
DIT:	0.25 s
Band width:	0.5 nm
Wavelength range:	210 nm to 180 nm
Data pitch:	0.05 nm
Scan speed:	100 nm/min
Accumulations:	1

Save the file as “04 Ammonia Vapour.jws”. Note the maximum HT value for the peaks. The HT maxima should be at 208.6 nm, 204.8 nm, 201.1 nm, 197.5 nm, 194.0 nm, 190.6 nm and 187.2 nm within 0.8 nm. Select Processing > Common Options > Channel Conversion. Select Channel 2, click Add > OK. Select Processing > Peak Processing > Peak Find. Save the file as “04 Ammonia Vapour Peaks.jws”. If the wavelength accuracy is not within specification, ensure the shift is constant across the wavelength range by checking with ACS as below. If the shift is constant recalibrate the spectrum accordingly. If it is not consult Professor Rodger who will call in an engineer.

5. *To determine the scaling factor for the CD intensity proceed as follows:*

5.1 Measure a spectrum of 0.06% (w/v) aqueous ammonium d-10-camphor sulfonate (ACS) in a 1 cm pathlength cuvette. Subtract a water baseline run with the same cuvette and parameters.

The ACS sample may be stored in a sealed glass container in the refrigerator. The instrument parameters are as follows:

Channels:	CD/HT/Abs
DIT:	0.5 s
Band width:	1 nm
Wavelength range:	350 nm to 250 nm
Data pitch:	0.05 nm
Scan speed:	50 nm/min
Accumulations:	1

Save the file as “05 0.06% ACS 1 cm cuvette.jws”. The wavelength and intensity of the CD peak should be 190.4 ± 1 mdeg at 291.0 nm. If the intensity is not within stated limits use an independently made ACS standard and repeat the calibration test. If the value is reproducible, all subsequent data should be scaled to bring the intensity to the correct value.

5.2 Test the performance of the instrument below 200 nm by placing the above ACS solution in a 1 mm cuvette and collecting a spectrum from 350 nm to 180 nm. Save the data as “05 0.06% ACS 1 mm cuvette.jws”. The ratio of the maxima of the 291.0 nm peak and the 193 nm should be 1:2.1.

5.3 Measure a spectrum of all three Na[Co(EDDS)] standards (blank, RR and SS) in the supplied Na[Co(EDDS)] calibration kit using the following parameters:

Channels:	CD/HT/Abs
DIT:	1 s
Band width:	1 nm
Wavelength range:	350 nm to 190 nm
Data pitch:	0.5 nm
Scan speed:	100 nm/min
Accumulations:	4

Save the files as “06 NaCoEDDS Blank/RR/SS.jws” Subtract the blank from the RR and SS samples. The wavelength and intensity of the main peak should be +50 ± 1 mdeg at 211.5 nm

for the SS and –50 ± 1 mdeg for the RR sample. Record the values in the instrument log as well as in the laboratory book for this assay.

6. *To measure the sample CD far-UV spectra proceed as follows:*

6.1 Remove the samples and appropriate buffers from the cold store and allow to reach room temperature prior to analysis.

6.2 Follow SOP28: *Jasco J-715 circular dichroism spectropolarimeter wavelength scans* modified to run with 10 L/min N_2 flow and to account for revised software.

6.3 Determine the required path length for the protein backbone region spectrum (from 260 nm to 190 nm) on the basis that a 1 mm cuvette probably requires a 0.1 mg/mL protein solution.

6.4 Clean the cuvette at each stage as follows. Rinse it well at least three times with de-ionised water followed by ethanol or acetone. Dry with a tissue, wipe with a lens cloth and remove any fibres with a nitrogen line prior to filling and reassembly for the demountable cuvettes. Dry with nitrogen for the non-demountable cuvettes. If the cuvette shows traces of protein residue, wash well with detergent, rinse with de-ionised water. If the residue still remains, place cuvette in a solution of 6 M nitric acid or Hellmanex. Allow the quartz to stand for ten minutes or longer if necessary before removing and rinsing thoroughly with water.

6.5 Fill the clean cuvette of the appropriate path length with high purity (18 MΩ) water and run a water spectrum on the J-715 following the instrument SOP and using the following parameter set: 0.2 nm (or other choice) data pitch, 100 nm/min, 8 accumulations (or the number required for satisfactory signal to noise levels), 1 s response, 1.0 nm band width. Save the data both in Jasco file format and as a text file. This spectrum is solely to check that the cuvette has been properly cleaned and does not have any protein residues stuck to it (as shown by the absence of a protein *CD* spectrum).

6.6 If the cuvette is not clean, repeat 6.4 and 6.5 until it is.

6.7 Repeat steps 6.4 and 6.5 but filling the cuvette with the buffer as supplied. This spectrum will henceforth be referred to as the baseline.

6.8 Repeat steps 6.3 and 6.4 but filling the cuvette with the protein reference standard or test sample(s).

7. *To measure the sample CD near-UV spectra proceed as follows:*

7.1 Remove the samples and appropriate buffers from the cold store and allow to reach room temperature prior to analysis.

7.2 Follow SOP28: *Jasco J-715 circular dichroism spectropolarimeter wavelength scans* modified as required for new software and collect data from 320 nm to 250 nm.

7.3 Determine the required path length for the protein 'aromatic' region spectrum (from 320 nm to 250 nm) on the basis that a 1 cm cuvette probably requires a 1 mg/mL protein solution. This value varies significantly from protein to protein depending in particular on the percentage of tryptophans present. The ideal sample will have an $A_{280\ nm}$ of approximately 1.

7.4 Clean the cuvette at each stage as follows. Rinse it well at least three times with de-ionised water followed by ethanol or acetone. Dry with nitrogen gas or a tissue, wipe with a lens cloth and remove any fibres. If the cuvette shows traces of protein residue, wash well with detergent, rinse with de-ionised water. If the residue still remains, place cuvette in a solution of 6 M nitric acid of Hellmanex. Allow the quartz to stand for ten minutes or longer if necessary before removing and rinsing thoroughly with water and drying as above.

7.5 Fill the clean cuvette of the appropriate path length with high purity (18 MΩ) water and run a water spectrum on the J-715 following the instrument SOP and using the following parameter set: 0.5 nm data pitch, 100 nm/min, 8 accumulations (in the first instance), 1 s response, 1.0 nm band width, and 20 mdeg. Save the data both in Jasco file format and as a text file. This spectrum is solely to check that the cuvette has been properly cleaned and does not have any protein residues stuck to it (as shown by the absence of a protein *CD* spectrum).

7.6 If the cuvette is not clean, repeat 7.4 and 7.5 until it is.

7.7 Repeat steps 7.4 and 7.5 but filling the cuvette with the buffer as supplied. This spectrum will henceforth be referred to as the baseline.

7.8 Repeat steps 7.3 and 7.4 but filling the cuvette with the protein reference standard or test sample(s).

8. *Preliminary data analysis*

8.1 Either within the Jasco software or a spreadsheet program such as Excel or a graph plotting one such as Kaleidagraph subtract the baseline spectrum from each sample spectrum. If necessary zero the spectra in the 250 – 260 nm region for the far UV spectra and in the 310 nm to 320 nm region for the near UV spectra. Plot the resulting spectrum. If this is done within the Jasco software save the final spectrum as a text file. Filter the data so that it is in 1 nm increments.

8.2 If the molar amino acid residue concentration is known, plot the spectra as $\Delta\varepsilon$ by first converting the *CD* spectra from millidegrees to absorbance units by dividing by 32980 and then converting from absorbance units to $\Delta\varepsilon$ using the Beer Lamber Law ($A = \varepsilon\, c\, l$, where c is concentration of amino acids in mol dm^{-3} for the far-UV and molar protein concentration for near UV, and l is path length in cm).

9. *CD structure fitting data analysis for far UV spectra*

9.1 For each sample that has been measured for which *CD* structure fitting is required, take the text file for the baseline subtracted and zeroed spectrum and edit it in Excel or another piece of software to produce the data in the following form. One title line containing anything followed by 71 lines (assuming fitting is being undertaken from 260 nm to 190 nm) of numbers corresponding to the *CD* spectrum in units of moles of (amino acids)$^{-1}$ dm^3 cm^{-1} with only two decimal places. This will require taking the second column of numbers from the original text file dividing each entry by (l*32980) (to convert from mdeg to absorbance units and from l cm to 1 cm pathlength) and also dividing by the molar concentration of

amino acid residues in the sample. For example, if the protein sample is 18 mg/mL, then the molar concentration is (18×(# of residues per molecule))/ (MW of protein). If you have more than one data set, the second set starts on the line directly below the first.

9.2 Ensure that the CDsstr program and the required associated files are located in a directory on the C drive of the computer you are using. The files include: procd190.tst; readme.cd; secstr.dta; bascd.dta; Cdsstr.exe. procd190.tst is a data file that can be used to test the program; it has three data sets in it.

9.3 To run the program, proceed as follows. Delete, rename, or move any file with a .out filename extension remaining in the cdsstr folder.

9.4 Delete any previously used file named proCD.dta unless you wish to use it in the current run.

9.5 If it is not already available, prepare an input file called procd.dta containing the *CD* data of the protein(s) to be analyzed. Save the file as c:\cdsstr\proCD.dta in the required format.

9.6 Begin the analysis by opening a DOS window within windows. Type 'c:'. Then type 'cd\cdsstr' at the command prompt. Type 'CDsstr' to run the program assuming the program is in the top directory.

9.7 Enter values for the program variables as prompted. NbasCD = 22; Nwave = 71; Npro = number of data sets in procd.dta; ncomb = 100; icombf = 100000.

9.8 When the command prompt reappears, view, print and record in the laboratory book the results of the analysis by inspecting the output files anal.out and reconCD.out.

10. *Sample analysis*

The samples should be analyzed in the following order:

Current analytical reference standard (if available)

Test samples

Current Analytical reference standard (if available)

11. *Acceptance Criteria*

11.1 All instrument calibration criteria must be met prior to assay of first samples.

11.2 The initial and final spectra if collected on a standard sample should be comparable with each other, and if appropriate, comparable with previous analyses of the reference standard.

12. *Report preparation*

Prepare the report to include the following: assay acceptance criteria, data, path length used for analysis, sample, buffer and reference IDs, the final spectra from the samples, $\Delta\varepsilon$ plots if required, a numerical table of the *CD* values against wavelength for each sample, and the results of the protein fitting program. Results are to be reported relative to the current analytical reference standard where possible. The results must be checked and countersigned prior to release.